Herausgeber

T. LÄNGLE
M. HEIZMANN

FORUM BILDVERARBEITUNG 2026
IMAGE PROCESSING FORUM 2026

KIT Scientific Publishing

M. Heizmann | T. Längle (Hrsg.)

FORUM BILDVERARBEITUNG 2026
IMAGE PROCESSING FORUM 2026

FORUM BILDVERARBEITUNG 2026
IMAGE PROCESSING FORUM 2026

Herausgegeben von
M. Heizmann und T. Längle

KIT Scientific Publishing

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark
of Karlsruhe Institute of Technology.
Reprint using the book cover is not allowed.

www.bibliothek.kit.edu/ksp.php | E-Mail: info@ksp.kit.edu | Shop: www.ksp.kit.edu



This document – excluding parts marked otherwise, the cover, pictures and graphs – is licensed under a Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0): https://creativecommons.org/licenses/by-sa/4.0/deed.en



The cover page is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0): https://creativecommons.org/licenses/by-nd/4.0/deed.en

Print on Demand 2026 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 2510-7224
ISBN 978-3-7315-1502-9
DOI 10.5445/KSP/1000196989

Vorwort

Bildverarbeitung spielt in vielen Bereichen der Technik und des Alltags eine wichtige Rolle zur Informationserfassung. Sie ist eine etablierte Technologie u. a. in der Mess- und Automatisierungstechnik, der Robotik, der Fahrzeugtechnik und der Unterhaltungselektronik. Wichtige Vorteile von Bildverarbeitungssystemen gegenüber anderen sensorischen Prinzipien bestehen u. a. darin, dass Bilder berührungslos gewonnen werden können und dass Bildsensoren inzwischen vergleichsweise günstig sind. Besonders spannend an der Bildverarbeitung ist aber, dass die Sensorik weitgehend dem menschlichen Leitsinn – dem Sehen – entspricht, aber nicht an die Beschränkungen des Menschen gebunden ist. Dies betrifft etwa die nutzbaren Spektralbereiche, die quantitative Interpretierbarkeit der Bilder, die gleichbleibende Aufmerksamkeit und Reproduzierbarkeit oder die Möglichkeit zur Erfassung hochdynamischer Prozesse. Auch wenn die Bildverarbeitung als Teil mehrerer Fachdisziplinen – u. a. Mess- und Automatisierungstechnik, Systemtheorie, Mathematik, Informatik, Optik, Lichttechnik, Mikrosystemtechnik – eine gewisse Reife erreicht hat, gibt es immer noch spannende neue Erkenntnisse. Gerade die Bildverarbeitung profitiert enorm von neuen Technologien wie z. B. dem maschinellen Lernen und der künstlichen Intelligenz oder neuen Sensorprinzipien wie etwa Event-based Vision.

Das »Forum Bildverarbeitung / Image Processing Forum« hat das Ziel, über solche aktuellen Trends und neuartige Lösungen in der Bildverarbeitung zu berichten und zum fachlichen Austausch zwischen Wissenschaft und Anwendung beizutragen. Es findet in jedem zweiten Jahr seit 2010 statt und wird inzwischen gemeinsam vom Geschäftsfeld Advanced Sensing des Fraunhofer-Instituts für Optronik, Systemtechnik und Bildverarbeitung IOSB und dem Institut für Industrielle Informationstechnik IIIT des Karlsruher Instituts für Technologie KIT organisiert. Dem Aufruf zur Einreichung von Beiträgen sind erfreulich viele Autoren gefolgt. Aus den Einreichungen konnte der Programmausschuss nach einer eingehenden Begutachtung hochwer-

tige Beiträge auswählen und den Schwerpunkten

- Bildgewinnung,
- Bildverarbeitung und Messtechnik,
- Robotik und Fahrzeuge,
- Maschinelles Lernen und künstliche Intelligenz, sowie
- Qualitätssicherung

zuordnen. Zusätzlich wurden in mehreren Tutorials Einblicke in aktuelle Forschungsansätze gegeben. Zu den Beiträgen wurden Aufsätze erstellt, die im vorliegenden Tagungsband enthalten sind. Wir danken den Autoren, den Mitgliedern des Programmausschusses für die Ansprache von Autoren, ihre wertvolle Expertise bei der Begutachtung der Einreichungen und allen, die durch ihre Anwesenheit zum Gelingen des »Forums Bildverarbeitung / Image Processing Forum« beitragen. Für die Organisation der Veranstaltung und die technische Unterstützung bei der Erstellung des Tagungsbands bedanken wir uns bei Lukas Dippon, Jürgen Hock und Britta Ost.

November 2026

Thomas Längle
Michael Heizmann

Tagungsleitung

Prof. Dr.-Ing. T. Längle	Fraunhofer IOSB Karlsruhe
Prof. Dr.-Ing. M. Heizmann	KIT IIIT

Mitglieder

Prof. Dr. C. Bach	NTB HS Buchs
Dr.-Ing. S. Bauer	Ubicept Boston (USA)
Dr.-Ing. D. Berndt	Fraunhofer IFF
Prof. Dr.-Ing. J. Beyerer	Fraunhofer IOSB
Prof. Dr. A. Braun	HS Düsseldorf-FEI
Dr. rer. nat. J. Eggert	Honda RI Europe
Dr. M. Glitzner	MVTec Software
Dr. T. Haist	ITO Uni Stuttgart
Prof. Dr. A. Heinrich	HTW Aalen
Prof. Dr. M. Huber	Fraunhofer IPA
Prof. Dr. B. Jähne	Heidelberg
M. Sc. C. Kludt	Fraunhofer IOSB
Dipl.-Ing. M. Maurer	VITRONIC
Dr. J. Meyer	Fraunhofer IOSB
Dr. M. Overdick	SICK AG
Dr.-Ing. M. Petz	TU Braunschweig
Prof. Dr. F. Salazar	Uni Madrid (Spanien)
Dipl.-Ing. M. Stelzl	MSTVision
Prof. Dr. R. Stiefelhagen	KIT SZS
Prof. Dr.-Ing. C. Stiller	KIT MRT
Prof. Dr.-Ing. M. Ulrich	KIT IPF
Prof. Dr.-Ing. S. Werling	DHBW Mannheim
Prof. Dr.-Ing. V. Willert	TU Darmstadt

Herausgeber

T. Längle und M. Heizmann

Contents

Vorwort i

Bildgewinnung

A new measurement technique for the image sensor MTF and its significance for modern image sensors 1
B. Jähne

Unified characterization of area-scan, event-based, and SPAD cameras using EMVA1288-inspired quality metrics 11
A. Manakov and B. Jähne

Multipoints - Eine Übersicht 23
T. Haist, N. Rustom, M. Tscherpel, A. Fuhrmann und S. Reichelt

Comparative analysis of integral transform-based single- and few-shot fringe image processing algorithms 35
Ö. Atmaca, W. Hasnaoui, M. Ebrahimnejad, and A. Herkommer

Maschinelles Lernen und künstliche Intelligenz

Kombination der Stärken von Simulation und generativer KI zur synthetischen Bildgenerierung 47
A. Sielemann, M. Roschani, J. Ziehn und J. Beyerer

Precise monocular ball localization in 4K through agentically generated synthetic training data 59
T. Pruß, J. Breuling, T. Dai, and R. I. Leine

Synthetic training data for YOLO-based optical 3D inspection of fork welds in manufacturing 71
J. Zender, S. Maier, and M. Layh

TorchGCvM: Fast Cramér-von-Mises distance for PyTorch 83
L. Dippon and J. Meyer

Uncertainty-guided dynamic routing for cost-efficient image classification 91
E. Rajoo, S. Landgraf, and M. Ulrich

Evaluating multi-view anomaly detection for inspection of reusable containers 103
V. Wahyudi and M. Storath

Cascaded outlier rejection for object-specific monocular time-to-collision estimation 115
M. Iorpenda, A. Sabotinov, and V. Willert

Beschreibung großer Bilddatensätze in natürlicher Sprache mit Vision-Language Models für das automatisierte Fahren 127
J. Truetsch, F. Bieder und C. Stiller

Robotik und Fahrzeuge

Optical quality and computer vision performance: Emerging new metrics 139
A. Braun

Non-contact brain-computer interface based on pupillary light reflex with extended frequency range* 151
N.O. Rohweder, M. Alizadeh, and C. Rembe

Laser-tracker-assisted camera-to-robot calibration for mobile robots 163
J. A. Rudolph, Ö. Kandemir, and M. Ulrich

Bildverarbeitung und Messtechnik

Single-shot deflectometry using an inverse crossed fringe pattern . 175
C. Kludt and J. Beyerer

LineAP: A segment-based average precision metric for visual rail track detection 187
A. Bätz and M. Lauer

On-camera person localization with omlox integration for safety applications 199
J. Ahlvers et al.

Autofocus-based surface reconstruction in digital holography 211
H. Sasikumar, B. Mukundakumar, and S. Matsuura

Tutorials

Perspektive als 3D Messverfahren: Ein Überblick 223
T. Haist, N. Rustom und S. Reichelt

Tutorial zur Hand-Auge-Kalibrierung 235
B. Erdnüß

A brief overview of quantum and quantum-inspired approaches for image segmentation 247
M. Gogolin, Y. Werner, A. Geng, P. Lukowicz, A. Moghiseh, and M. Kiefer-Emmanouilidis

Qualitätssicherung

Detection of surface defects on free-form components 259
D. Sopauschke, T. Dunker, and E. Trostmann

STEP-ID: Computer vision-based part identification for post-production automation in additive manufacturing 269
M. G. Schneider, L. Metz, O. Werger, R. Seidel, and A. Grabmaier

Zeitaufgelöste Brechungsindexentwicklung in DLP-gehärteten Photopolymeren für mikrooptische Anwendungen 283
A. Heinrich

A new measurement technique for the image sensor MTF and its significance for modern image sensors

Bernd Jähne

IWR, Heidelberg University
Im Neuenheimer Fels 205, 69120 Heidelberg, GERMANY

Abstract This paper investigates using the Fresnel zone plate pattern to measure the two-dimensional modulation transfer function (2-D MTF). This pattern is widely used to evaluate filter performance in image processing. For MTF measurements, the pattern's significant advantage is that it maps the Fourier domain back to the spatial domain. Thus, the wavenumber is known and MTF measurements can be extended beyond the Nyquist limit. The MTF technique and computation are described and verified by simulation.

Keywords Image sensor, MTF, EMVA 1288 standard

1 Introduction

The resolution of image sensors without lenses used to be of interest only to specialists in this field for several reasons. Firstly, the lens was the main factor restricting image resolution. Secondly, image sensors simply performed a straightforward spatial averaging of the light-sensitive image of a pixel. However, with progress in both lens design and image sensor technology, this is gradually changing. The spatial resolution of lenses has improved significantly, with lenses now being diffraction-limited at smaller aperture numbers. Furthermore, image sensors and cameras are increasingly performing embedded processing to improve image quality. Common examples include noise suppression and de-bayering for colour images. These processing steps alter the spatial resolution and signal-to-noise ratio (SNR) of images.

Standards, such as the EMVA 1288 standard [1, 2], that characterize machine vision cameras as components without lenses currently do not include any measurements of how the spatial sampling performed by an image sensor and subsequent preprocessing in the camera influence or limits resolution. This is a serious limitation of current parameters that characterize the quality of image sensor signals. Even if preprocessing improves the image sensor signal-to-noise ratio (SNR), it is unclear to what extent the spatial resolution is limited. Therefore, it is important to measure and compare the spatial resolution of image sensors without lenses, as well as the spatial resolution of raw and preprocessed image sensor data.

The two-dimensional point spread function (PSF) and modulation transfer function (MTF) accurately measure the spatial resolution of image sensors. Techniques commonly used to measure MTF project patterns onto image sensors using interferometry [3], laser speckle patterns [4], or use a high-resolution microscopic lens to image slanted edges [5] or the Siemens star onto the image sensor plane. A significant disadvantage of interferometric and laser techniques is that they can only measure the MTF for a single wavelength. The other techniques require significant effort to compute the two-dimensional (2-D) MTF.

This paper investigates a new method of measuring the modulation transfer function (MTF) of image sensors. This method uses a high-resolution microscopic lens to project an image onto the sensor. Section 2 describes the pattern and its advantages for MTF measurements, and Section 3 explains how to compute the MTF. Section 4 demonstrates the new technique using simulations. The paper concludes with an overview of how these measurements can be performed in practice.

2 The Fresnel Zone Plate as a New Pattern for MTF Measurements

The search for a suitable pattern to project onto an image sensor was guided by the following question: What is the easiest way to determine the two-dimensional modulation transfer function (MTF)? This brought the attention to a pattern extensively used to evaluate filter performance in image processing [6]. It consists of concentric rings,

where the gray value changes sinusoidally in the radial direction, calculated with

$$g(\boldsymbol{x}) = a_0 \sin\left(\frac{\pi\kappa_m \left|\boldsymbol{x}\right|^2}{r_m}\right)\left[\frac{1}{2}\tanh\left(\frac{r_m - \left|\boldsymbol{x}\right|}{w}\right) + \frac{1}{2}\right]. \tag{1}$$

Where a_0 is the amplitude of the pattern, r_m is the maximum radius of the pattern, κ_m is the maximum wavenumber at the edge and w is the width of the drop to zero at the maximum radius. The wavenumber increases linearly from zero in the center to κ_m at the edge, as can be calculated by derivation of Eq. 1:

$$\boldsymbol{\kappa} = \frac{\kappa_m \boldsymbol{x}}{r_m}. \tag{2}$$

The vector $\boldsymbol{x}$ is the vector from the center of the pattern to the corresponding position. In this way it is easily possible to calculate the wavenumber vector at each position in the test pattern.

In optics, the pattern is known as Fresnel zone plate. In the ring test pattern, structures occur in all directions and all wavenumbers up to a selectable maximum.

Furthermore, the wavenumbers are represented in the image as in two-dimensional Fourier space with the center of the coordinate system in the middle of the pattern. Therefore, the optical transfer function (OTF) — i.e., the change in amplitude and phase shift — is directly visible when a digital image of this pattern is acquired by an image sensor.

For MTF measurements, this pattern has the significant advantage of mapping the Fourier domain back to the spatial domain. Thus, the wavenumber is known, and extending the MTF measurements beyond the Nyquist limit is easily possible. Figure 1 illustrates how the patterns look like if the wavenumber is extended beyond the Nyquist limit.

As we approach the limit, the patterns become distorted because there are slightly more than two samples per period. Then, the sampled amplitude changes with the phase. Sampling the pattern at the maximum and minimum captures the full amplitude. With a 90° phase shift, sampling occurs at the zero crossings, and the pattern becomes invisible.

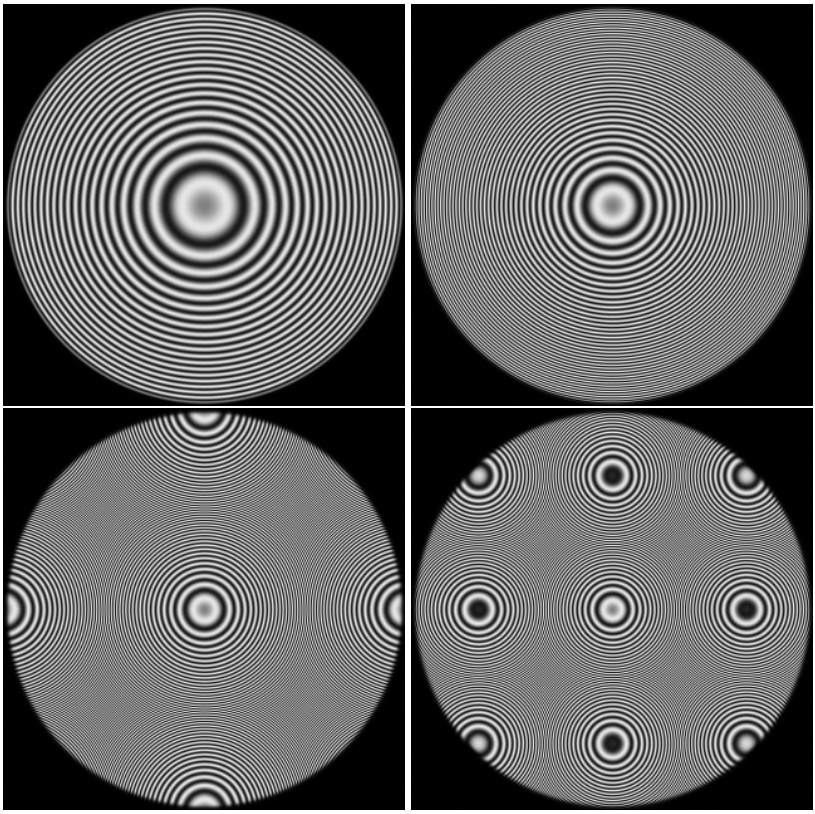


Figure 1: Fresnel zone plate as a MTF test pattern with maximum wavenumbers relative to the Nyquist wavenumber: 0.5 (top left), 1.0 (top right), 2.0 (bottom left), and 3.0 (bottom right).

Beyond the Nyquist limit, the Fourier space is replicated. This is a consequence of the sampling theorem. This leads to basic multiplicity. It is unclear whether the sampled wavenumber is the actual wavenumber or the wavenumber plus a vector of the reciprocal grid. However, the true wavenumber can be determined from Eq. (2). Thus, the analysis of the MTF can be extended beyond the Nyquist limit.

3 Computation of the MTF with the Phase-Shift Technique

The MTF represents the amplitude of the periodic pattern. However, the sampled amplitude depends on the phase. As shown in the previous section, it varies from zero to the true amplitude. One solution to this problem is to compute the monogenic signal using a Riesz transform [**?**]. However, if the amplitudes are sampled close to the zero crossings they are small. This leads to noisy results.

A better way is to introduce a phase shift to the ring pattern

$$g(\boldsymbol{x}) = a_0 \sin\left(\frac{\pi\kappa_m \left|\boldsymbol{x}\right|^2}{r_m} + \Delta\varphi\right)\left[\frac{1}{2}\tanh\left(\frac{r_m - \left|\boldsymbol{x}\right|}{w}\right) + \frac{1}{2}\right] + b. \quad (3)$$

This equation also includes as an additional component an unknown background intensity, b. Therefore, there are three unknowns in total: background brightness (b), amplitude of modulation (a_0), and unknown phase shift ($\Delta\varphi$).

To determine the three unknowns of the pattern, it is necessary to project at least three of the same patterns onto the objects with equidistant phase shifts. We consider the general case of P equally sampled phase shifts within one period, i. e the interval from 0 to 2π:

$$\varphi_n = \frac{2\pi p}{P} \quad \text{with} \quad p = \{0, 1, \ldots, P-1\}. \quad (4)$$

This results in a sequence with P brightness values g_p at every point in the image. This means that we sample the periodic function equidistantly over a wavelength of P values. The three unknowns result directly from the fact that the signal in Eq. (3) is composed of a constant component b and a periodic one with the discrete wavenumber one. This allows us to directly apply the discrete Fourier transform. For the constant part, this results in

$$b = \frac{1}{P}\sum_{p=0}^{P-1} g_p \quad (5)$$

and for wavenumber one an amplitude and phase shift of

$$a_0 = \frac{2}{N}\left|\sum_{p=0}^{P-1} g_p \exp\left(\frac{2\pi \mathrm{i} p}{P}\right)\right|, \quad \Delta\varphi = \arg\left[\sum_{p=0}^{P-1} g_p \exp\left(\frac{2\pi \mathrm{i} p}{P}\right)\right]. \quad (6)$$

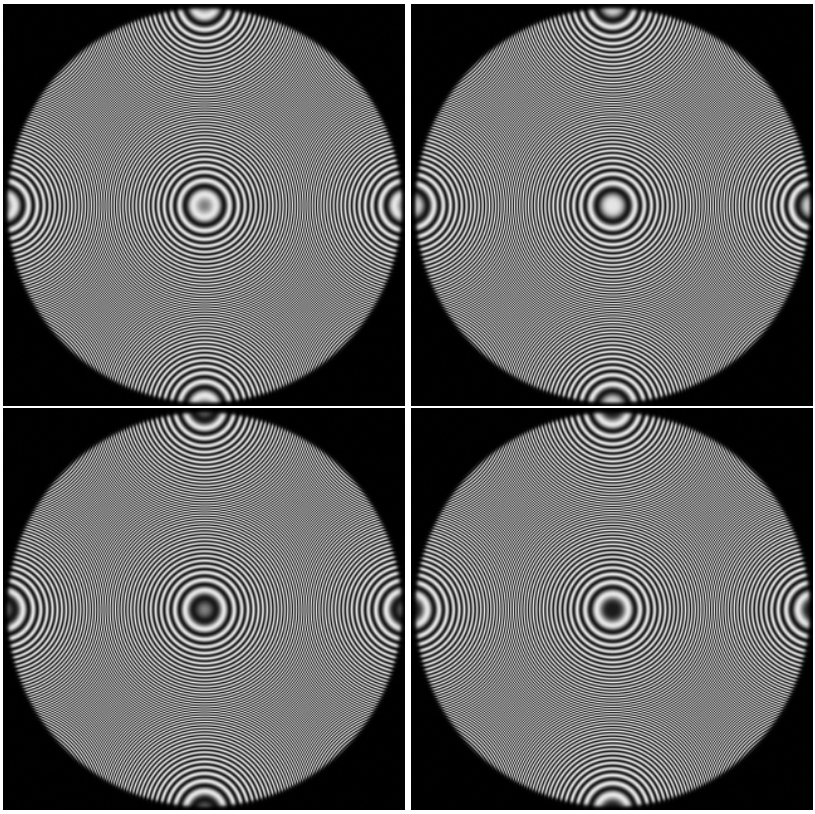


Figure 2: Fresnel zone plate as a MTF test pattern with maximum wavenumbers relative to the Nyquist wavenumber of 2 and phase shifts of 0°, 90°, 180°, and 270°.

Computing the amplitude and phase is particularly simple with four samples. In this case, the amplitude and phase position Eq. (6) simplify to

$$g_m = \frac{1}{2}\sqrt{(g_0 - g_2)^2 + (g_1 - g_3)^2} \quad \text{and} \quad \Delta\varphi = \arctan\frac{g_3 - g_1}{g_2 - g_0}. \tag{7}$$

Therefore, this sampling has been chosen. The resulting four ring patterns with wavenumbers up to twice the Nyquist wavenumber are shown in Fig. 2

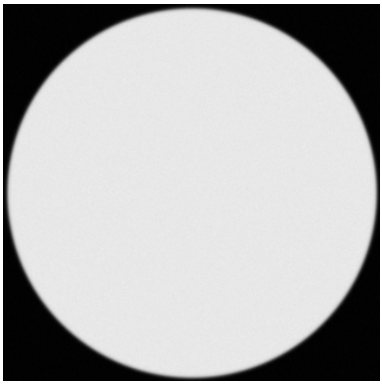


Figure 3: Estimate of the amplitude of an MTF test pattern with twice the maximum wavenumbers relative to the Nyquist wavenumber. Gaussian noise has been added to all four phase-shifted patterns in Fig. 2. The standard deviation of this noise is 1% of the amplitude of the periodic pattern.

The computational method described here is not new. It has been used for other image acquisition techniques. These include the phase shift technique for active triangulation with fringe images and time-of-flight imaging with periodically modulated light [6].

4 Demonstration by Simulation

This section verifies the correctness of the amplitude computation, and thus the MTF, by simulation. This step is necessary because the pattern does not have a constant wavenumber, which is what the equations were derived for. Computing the amplitude from the four images in Fig. 2 should result in a constant value, even though the wavenumbers are up to two times higher than the limiting Nyquist wavenumber, at which point the periodic pattern is sampled only twice per wavelength. This is an effective way to test whether the MTF analysis can be applied beyond the Nyquist wavelength. For a realistic simulation Gaussian noise with a standard deviation of 1% was added to all four test patterns in Fig. 2.

Figure 3 verifies that a constant amplitude is indeed computed at the Nyquist wavenumber and beyond. The histogram in Fig. 4 shows

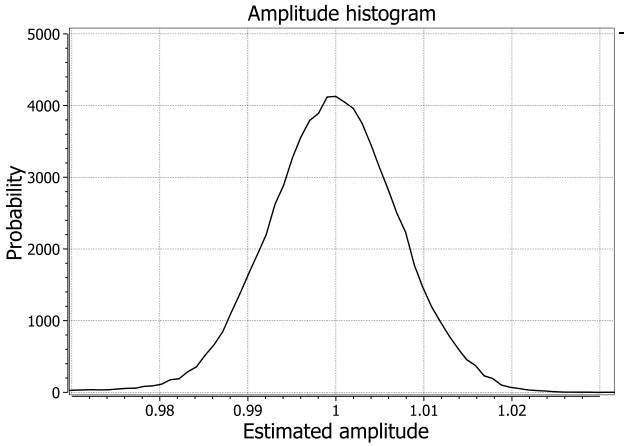


Figure 4: Histogram of the amplitude relative to the exact value from the amplitude image in Fig. 3.

shows a symmetric distribution around the correct value, with a standard deviation of less than 1% of the amplitude. This verifies that the MTF can be computed accurately enough, even with noisy test images. The standard deviation of the MTF value is smaller than the standard deviation of the test patterns relative to their amplitudes.

5 Conclusion and Outlook

This paper establishes a new method for measuring the MTF of image sensors. Using four phase-shifted ring patterns, we can measure two-dimensional MTF beyond the Nyquist limit. This is especially important for image sensors. When the MTF is non-zero beyond the Nyquist limit, Moiré patterns appear in the image.

The practical application of the technique requires two additional steps. First, corrections are necessary for the MTF of the lens when the pattern is projected by a microscopic lens onto the image sensor. An interesting question is how small the pixel size can be and still allow the technique to work. Second, the monitor used to display the patterns will not have a linear tone value curve. Correction will be necessary

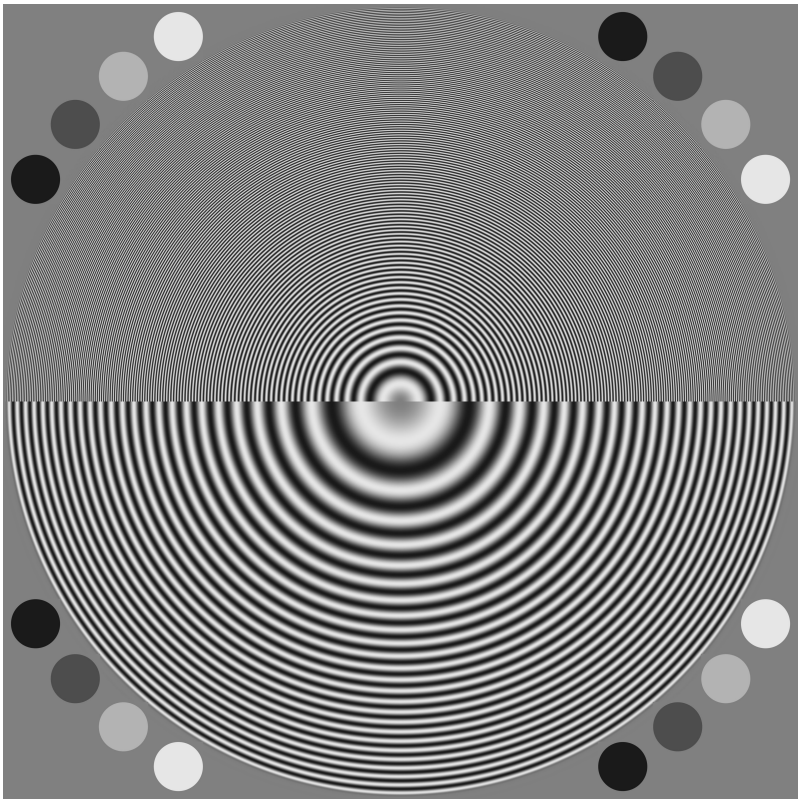


Figure 5: Test pattern for MTF measurements with additional dots of defined gray values to linearize the tone curve of the monitor.

for this effect. This can be performed easily if the test pattern contains areas with defined gray values as shown in Fig. 5.

The next step will be to set the whole system up and to perform systematic measurements with various industrial cameras. A special focus will be placed on comparing the MTF of raw image data and data that has been processed to improve image quality or demosaic color and multispectral image sensors.

Acknowledgements

The author gratefully acknowledges financial support for this research through his senior professorship at Heidelberg University.

References

1. EMVA 1288 Working Group, "EMVA Standard 1288 - standard for characterization of image sensors and cameras, release 4.0 linear," European Machine Vision Association, open standard, 2021. [Online]. Available: https://www.emva.org/standards-technology/emva-1288/emva-standard-1288-downloads-2/
2. B. Jähne, "Release 4.1 of the EMVA standard 1288: a universal concept to characterize modern image sensors," in *Forum Bildverarbeitung*, M. Heizmann and T. Längle, Eds. KIT Scientific Publishing, 2024, pp. 13–22.
3. T. Georgiev, J. Gille, A. Sun, L. Baev, and T. Battula, "Interferometric measurement of sensor MTF and crosstalk," in *Proc. Int. Symp. on Electronic Imaging: Digital Photography and Mobile Imaging XIII*, vol. 29, no. 15. Society for Imaging Science and Technology, 2017, pp. 52–57.
4. X. Chen, N. George, G. Agranov, C. Liu, and B. Gravelle, "Sensor modulation transfer function measurement using band-limited laser speckle," *Optics Express*, vol. 16, no. 24, p. 20047, 2008.
5. M. Estribeau and P. Magnan, "Fast MTF measurement of CMOS imagers using iso 12333 slanted-edge methodology," in *Detectors and Associated Signal Processing*, J.-P. Chatard and P. N. J. Dennis, Eds., vol. 5251. SPIE, 2004, p. 243.
6. B. Jähne, *Digital Image Processing and Image Acquisition*. Cham: Springer, 2026.

Unified characterization of area-scan, event-based, and SPAD cameras using EMVA1288-inspired quality metrics

Alkhazur Manakov and Bernd Jähne

IWR, Heidelberg University,
Berliner Straße 43, 69120 Heidelberg

Abstract This work considers three classes of imaging sensor that exhibit fundamentally different noise characteristics, data representations and temporal behavior. This complicates direct performance comparison and application-driven selection. Building on the systems-theoretic foundation of the EMVA 1288 standard, we continue our work from earlier publications [1] and [2]. SPAD sensors are incorporated into the analysis, and a new metric for event-based vision is introduced that quantifies the minimum number of photons required to detect an event against a dark background: the minimum detectable event (MDE). This is equivalent to the absolute sensitivity threshold of standard cameras in the EMVA 1288 standard. We also investigate how the MDE depends on sensor temperature.

Keywords Sensor characterisation, event-based, SPAD, area-scan, performance, comparison

1 Introduction

This work considers three classes of imaging sensors—conventional area-scan cameras, event-based vision sensors, and single-photon avalanche diode (SPAD) imagers—that differ fundamentally in their acquisition principles, data representations, noise characteristics, and temporal behavior, thereby complicating direct performance comparison and application-driven selection. Area-scan cameras acquire intensity images at discrete frame rates, yielding dense spatial infor-

mation but imposing constraints on temporal resolution and bandwidth scalability. Their frame-based operation samples the visual scene as a sequence of synchronous "snapshots" at a predetermined rate, a paradigm that is intrinsically misaligned with the asynchronous, continuous-time nature of real-world dynamics. This mismatch forces a trade-off between information loss in the inter-frame blind time. The adoption of high frame rates results in more complex acquisition and processing hardware together with high-bandwidth interconnects. Moreover, each frame records data from all pixels irrespective of whether the underlying scene content has changed, so adjacent frames in dynamic scenes contain substantial redundancy. The acquisition and handling of such dispensable data increase transmission power dissipation, channel bandwidth requirements, memory footprint, and processing load, motivating bio-inspired, pixel-individual, frame-free approaches as potentially more efficient alternatives to traditional machine vision.

The most advanced bio-inspired vision sensors currently available [3] adopt a natural, event-driven, frame-free paradigm that captures transient changes in the visual scene. At each pixel, analog circuitry maintains a stored reference brightness level and continuously compares it with the instantaneous illumination. When the deviation between the current and reference brightness exceeds a predefined threshold, the pixel resets its reference level and emits an event: a discrete data packet comprising the pixel address, a precise timestamp, and the polarity (increase or decrease) of the brightness change. Certain implementations within this class additionally provide instantaneous measurements of the absolute illumination level [4]. Sensors operating according to this principle are referred to as *event-based sensors*.

Single-photon avalanche diode (SPAD) imagers operate in a photon-counting regime, where each pixel functions as a Geiger-mode avalanche diode capable of detecting individual photons with sub-nanosecond timing precision [5], [6]. Upon absorption of a photon in the high-field multiplication region, a self-sustaining avalanche is triggered, producing a macroscopic current pulse that marks the photon arrival time with picosecond-scale resolution. SPAD arrays are typically operated in a time-resolved mode enabling applications such as time-of-flight ranging, fluorescence lifetime imaging, and transient imaging [7]. The noise behavior of SPADs is dominated by

dark counts—thermally or tunneling-induced avalanches in the absence of light—afterpulsing, in which trapped carriers from a previous avalanche are released and trigger spurious secondary events, and inter-pixel optical or electrical crosstalk. To mitigate afterpulsing, a dead time (hold-off interval) of tens to hundreds of nanoseconds is enforced after each avalanche, during which the pixel is insensitive to incoming photons; this dead time limits the maximum count rate and introduces non-Poissonian statistics that must be accounted for in quantitative photon-counting applications [6]. Consequently, SPAD-based imagers offer extreme sensitivity and unparalleled temporal resolution in low-light and time-resolved scenarios. Their data are represented as time-tagged binary frames, rather than conventional intensity frames, and their performance is governed by figures of merit such as photon detection probability, dark count rate, afterpulsing probability, dead time, and timing jitter.

Comparing cameras equipped with event-based, SPAD, and conventional area-scan sensors for a given machine vision application remains challenging, as a unified methodology for cross-technology performance evaluation is currently lacking. A first step toward such a framework was undertaken by Manakov and Jähne [8], who formalized the fundamental operating principles of event-based sensors and extended the EMVA 1288 characterization standard to this domain. Building on this foundation, Manakov et al. [9] proposed an experimental setup, data acquisition protocol, and initial measurement results for event-based sensor characterization. Subsequently, Manakov et al. [1] introduced key performance indicators—including Event Contrast Detectability (ECD), latency, and Event Detection Efficiency (EDE)—that are applicable to both event-based and conventional area-scan sensors, thereby enabling more direct comparative assessments. Theoretical analysis and supporting experimental results for these metrics have been presented in [2].

In this work single-photon avalanche diode (SPAD) sensors with configurable bit depth are embedded into the event-based sensor behavior emulation framework introduced in [2]. The framework thus enables a unified comparison of event-based, area-scan, and SPAD devices.

We further introduce the concept of a minimum detectable contrast for all sensor types considered in this work, which is conceptually analogous to the absolute sensitivity threshold defined in EMVA 1288

for area-scan and SPAD sensors. In the event-based case, the idea is straightforward: we characterize sensor performance in complete darkness by analyzing how the event probability depends on the applied brightness-change amplitude.

The performance is compared against those of an idealized event-based sensor with unit quantum efficiency and no noise sources other than photon shot noise, allowing us to interpret the minimum detectable contrast as the contrast level at which the event probability—and thus the effective SNR—crosses a defined detection threshold, in direct analogy to the SNR = 1 criterion used for the absolute sensitivity threshold. We additionally investigate how the resulting minimum detectable contrast varies with sensor temperature.

2 Event-based sensor characterization

The event-based characterization method defines event probability as the ratio of event responses to repeated identical stimuli at fixed background irradiance and sensor settings [8], [10]. The characterization the measurement of how this probability varies with the irradiance step amplitude to form an S-curve. From each S-curve, key performance parameters are extracted: the contrast at a give event probability (for instance, 0.5), the slope at that point (indicating noise influence), and any vertical offset (revealing spontaneous events without temporal contrast) [1]. These S-curves are acquired over a wide range of background irradiances using an EMVA 1288-compliant setup, enabling a unified, quantitative description of an event-based sensor.

S-curve is a useful performance indicator of an event-based sensor, but to enable the comparison of contrast detection performance between the sensor types quantitative performance metrics were introduced by Manakov et al. [2]. Event Contrast Detectability (ECD) quantifies how large a relative irradiance change is required for a sensor to reliably detect a temporal contrast. It is defined from the S-curve as the inverse of the steepest slope at the 0.5 event-probability point, scaled by the irradiance at that point:

$$\mathrm{ECD} = \frac{\Delta E}{E_{0.5}} = \left(E_{0.5} \frac{\mathrm{d}S}{\mathrm{d}E} \right)^{-1}, \tag{1}$$

where $S(E)$ is the event probability versus irradiance step. A lower ECD means the sensor can detect smaller relative changes; for example, ECD ≈ 0.3 implies that roughly a third increase in irradiance is needed to trigger events with high probability. Because ECD can be derived analytically for linear area-scan sensors via their SNR, it provides a common metric to compare event-based and conventional cameras on equal footing [2].

Event Detection Efficiency (EDE) measures how close a real camera's contrast-detection performance comes to that of an ideal, noise-limited sensor with quantum efficiency 1 and no dark noise. It is defined as the ratio of the ideal ECD to the measured ECD at a given irradiance

$$\mathrm{EDE} = \frac{\mathrm{ECD}_{\mathrm{ideal}}}{\mathrm{ECD}} = \frac{2\sqrt{2\pi}}{\mathrm{ECD}} \sqrt{EAt_{\mathrm{exp}}}. \tag{2}$$

For a linear area-scan sensor:

$$\mathrm{EDE}_{\mathrm{lin}} = \frac{\eta \sqrt{EAt_{\mathrm{exp}}}}{\sqrt{\sigma_d^2 + \eta EAt_{\mathrm{exp}}}}, \tag{3}$$

and at high irradiance (negligible dark noise) EDE $\approx \sqrt{\eta}$. EDE thus captures the combined impact of quantum efficiency and noise on detectability, and when applied to event-based sensors (using latency in place of exposure time) it reveals how much room remains for technical improvement, especially in dark-noise performance at low light levels.

The emulation of event-based behavior by an area-scan sensor treats the camera as a temporal change detector that compares two consecutive frames and declares an event whenever the gray-value difference exceeds a fixed threshold. Assuming Gaussian temporal noise, the probability of detecting such a change as a function of the underlying signal difference forms an S-shaped curve analogous to the event-probability curves measured on true event-based sensors. This analytical model lets one compute the same contrast-detectability metrics used for event cameras directly from standard EMVA 1288 parameters such as SNR and exposure, providing a unified way to predict how a conventional sensor would behave in an event-driven mode without performing new measurements.

SPAD sensors can be treated as area-scan cameras by accumulating a sequence of binary (1-bit) frames and summing them to synthesize a

multi-bit intensity image. Each binary frame records only whether at least one photon was detected during a very short exposure. A single binary frame has extremely limited dynamic range. By summing M such frames per pixel, the sensor effectively performs photon counting. The resulting integer value can be interpreted as a grayscale level with a bit depth determined by the number of accumulated frames (for example, 2^{b-1} frames to approximate b-bit depth). SPAD arrays have negligible readout noise and can operate at very high frame rates, this frame-summation approach yields high-sensitivity, low-noise images that behave like a conventional area-scan camera. Owing to this property, SPAD sensors can be characterized using the same ECD and EDE metrics as conventional area-scan cameras. The controllable aggregation of binary SPAD frames provides a tunable trade-off between effective bit depth-governing the smallest detectable irradiance contrast-and the equivalent exposure time of the synthesized intensity image.

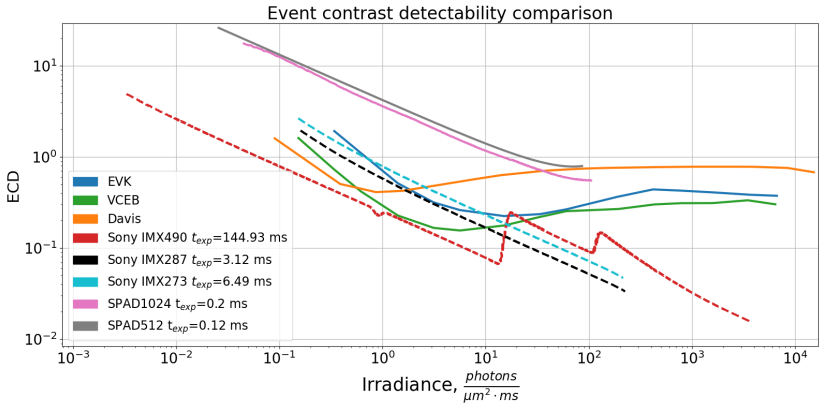


Figure 1: Event detectability comparison for area-scan, SPAD and event-based sensors for positive event polarity. SPAD1024 and SONY IMX490 are equipped with Bayer filters, for these sensors data of the green channel is presented.

ECD was computed for all devices under test as described in [2], and the results are shown in Figure 1. Lower ECD values correspond to superior contrast detection. Standard linear industrial area-scan sensors achieve lower ECD than event-based cameras at moderate and high irradiance, reflecting their ability to detect smaller relative irradiance

changes, but they do not match the dynamic range of event-based devices. The HDR camera based on the SONY IMX490 exceeds the dynamic range of all tested event-based sensors and also attains very low ECD at low irradiance. The SPAD sensors display comparatively poor contrast detectability, requiring larger irradiance differences between successive frames to register an event. This is partly attributable to their limited effective bit depth: SPAD512 and SPAD1024 were characterized with 6-bit and 7-bit reconstructed frames, respectively, versus 12-bit for the area-scan cameras, and SPAD1024 was measured under white light with a Bayer filter, reducing the number of photons reaching the active pixel area. The SPAD sensors support exposure-time sweeping, combining the low-ECD segments from multiple exposure settings enables near-optimal contrast detection over a broad irradiance range.

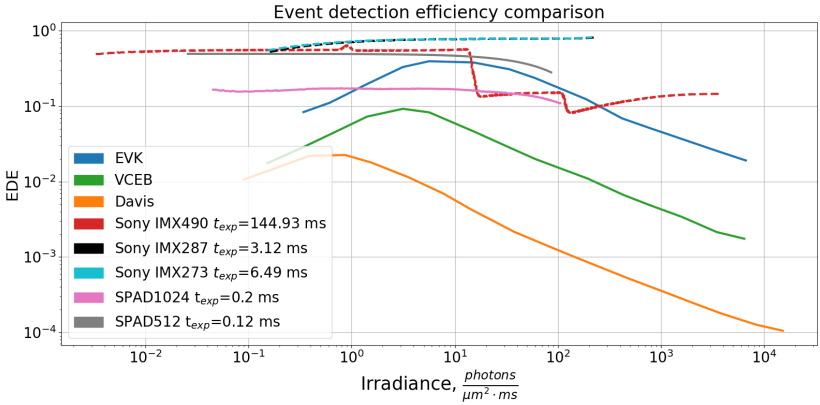


Figure 2: Event detectability comparison for area-scan, SPAD and event-based sensors for positive event polarity. SPAD1024 and SONY IMX490 are equipped with Bayer filters, for these sensors data of the green channel is presented.

The tested area-scan cameras operate close to this theoretical limit (EDE=1), reflecting their high quantum efficiency and low temporal dark noise (Figure 2). Among the event-based sensors, a clear progression is visible from the DAVIS device to the Prophesee Gen 3.1 (VCEB) and Gen 4.1 (EVK) cameras. The comparison in Figure 1 implicitly assumes a digital comparator applied to consecutive area-scan or SPAD frames to decide on event occurrence, whereas event-based

sensors employ an analog in-pixel comparator that enables faster detection but introduces additional noise. The observed decline in EDE at high irradiance for event-based devices is not detrimental, as it is accompanied by only a modest degradation in ECD in the same regime (Figure 1). SPAD sensors outperform older event-based models and reach efficiency levels comparable to the area-scan cameras under test. SPAD512 exhibits higher EDE than SPAD1024, primarily due to its superior quantum efficiency: it was characterized with green light near its spectral quantum efficiency peak and without Bayer filters, whereas SPAD1024 was measured under white light with a color filter array that transmits only a small fraction of photons to the active pixel area.

3 Minimum detectable event

The *minimum detectable event (MDE)* for an event-based camera is defined under complete darkness, where the sensor is stimulated with monotonically increasing irradiance steps and the event probability is measured for each stimulus amplitude. MDE is the irradiance level change at which the event probability first reaches 0.5, indicating the minimum detectable stimulus in the absence of any background illumination. This metric characterizes the sensor's intrinsic noise floor and its ability to generate events from near-zero light conditions.

MDE is measured by acquiring a family of S-curves under zero background illumination (Figure 3), i.e., in complete darkness. For each stimulus amplitude, the sensor is repeatedly stimulated with identical irradiance steps while the background remains at zero. The stimulus level corresponding to 0.5 event probability is then converted into an equivalent number of incident photons using the calibrated optical setup. This photon count at the 0.5 event-probability point constitutes the MDE of the event-based sensor.

The S-curves presented in Figure 3 were acquired for the VCEB and EVK event-based sensors at multiple operating temperatures and with various neutral-density (ND) filter combinations to span a wide irradiance range [2]. Across the tested conditions, the S-curves for both sensors are largely consistent, with the minimum detectable contrast (MDE) increasing only slightly as temperature rises; this indicates a modest reduction in sensitivity at higher temperatures but confirms

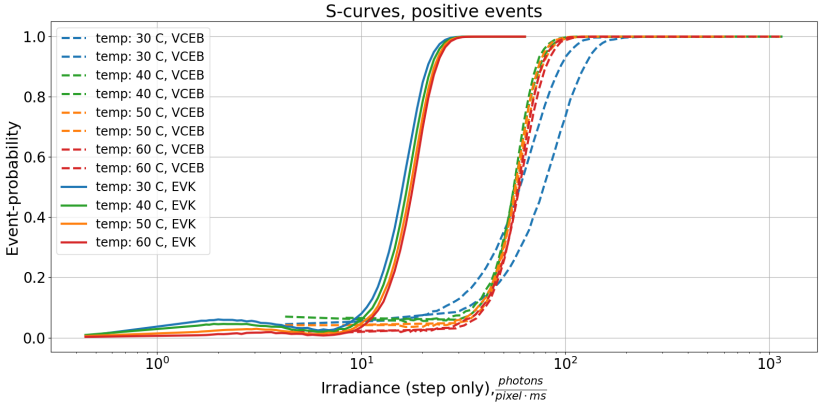


Figure 3: S-curves for EVK and VCEB acquired at different sensor temperatures.

that false-positive events in complete darkness remain negligible even under increased temperature operating conditions. One notable exception is the S-curve for the VCEB sensor at 30 degree Celsius, which deviates from the otherwise coherent family of curves. The origin of this inconsistency is currently unknown and may stem from a measurement artifact or point to an physical effect warranting further investigation.

Figure 4 shows that MDE increases only slightly with increasing sensor temperature. The data point for the VCEB sensor at 30 degrees has been omitted from the MDE–temperature dependency plot due to the aforementioned inconsistency in its S-curve (Figure 3). Surprisingly, the rate of false positives at an irradiance below the MDE threshold is higher at lower temperatures. Clearly, the event-based sensors incorporate an as yet unknown mechanism to suppress the typically increasing dark signal with temperature. The plot also reveals a clear generational difference: VCEB device exhibits a higher MDE than EVK sensor, meaning that generation 3 requires a larger stimulus in complete darkness to reach the 0.5 event-probability point. Conversely, the generation 4 sensor achieves the same event probability at lower stimulus.

The MDE calculation for SPAD sensors can be reformulated as determining the number of incident photons N to sensor with a given

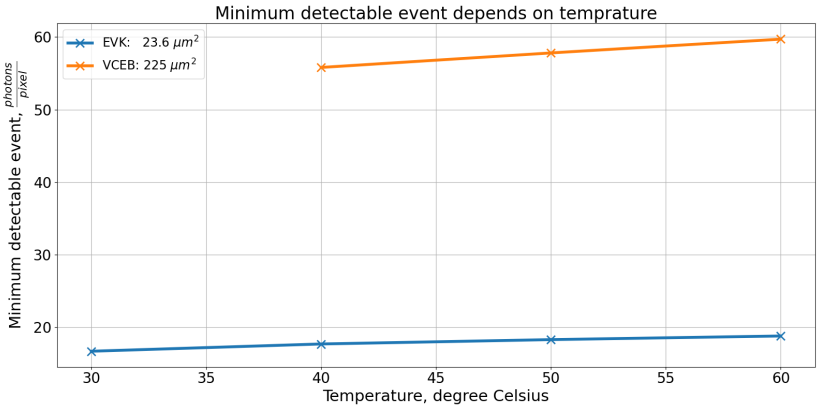


Figure 4: Minimum detectable event dependency on the temperature for EVK and VCEB.

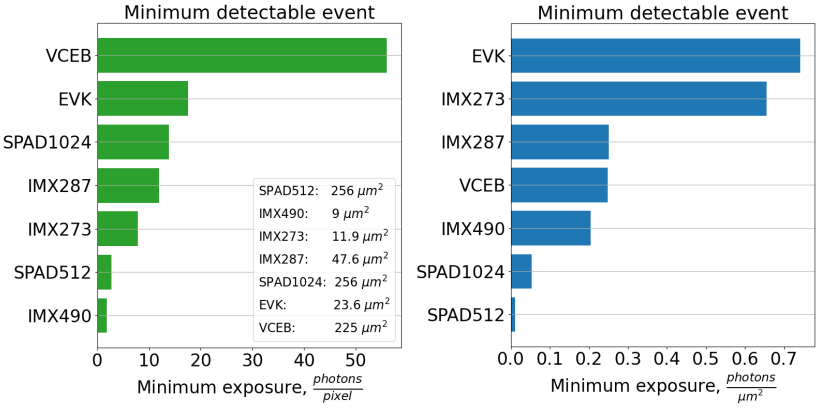


Figure 5: MDE for SPAD, event-based ans area-scan sensors. Left - in photons per pixel, Right - in photons per μm^2.

quantum efficiency η detects at least one photon with 0.5 probability. The probability of detecting a photon is given by Poisson statistic $P(1) = e^{-\mu}$, where $\mu = N \cdot \eta$. Thus, N required for detecting a photon with 0.5 probability is given by $N = ln(2)/\eta$. For area-scan sensors that emulate event-based behaviour, the contrast detection threshold defined by Manakov et al. cannot be applied in the case of zero background illumination. The required threshold for event detection is instead determined by dark noise. If the dark current noise distribution is approximated as Gaussian (see EMVA 2021), the threshold should be set to 2.34 standard deviations above the mean. This limits the false-positive event rate to 1% in the absence of any temporal contrast. Depending on the application, a different acceptable false-positive event rate may be chosen, in which case the corresponding threshold multiplier will change slightly. Figure 5 compares the MDEs. The ranking depends on whether the MDE is given in photons per pixel or photons per unit area (i. e. constant irradiation).

4 Conclusion and outlook

This work establishes a unified characterization framework for area-scan, event-based, and SPAD sensors by extending EMVA 1288-inspired metrics—Event Contrast Detectability (ECD) and Event Detection Efficiency (EDE)—to all three technologies, enabling direct, quantitative performance comparison. Area-scan sensor operate near the theoretical efficiency limit, while event-based sensors show clear generational progress, with Gen 4 devices surpassing Gen 3 in both contrast detectability and noise efficiency. SPAD imagers, when operated in frame-accumulation mode, integrate seamlessly into this framework and achieve performance comparable to conventional sensors despite their binary native output. The introduced MDE, defined as the irradiance yielding 0.5 event probability in darkness provides a common low-light sensitivity metric across all sensor types. MDE reveals modest temperature dependence and superior performance of newer event-based generations. This framework supports application-driven sensor selection and outlines a path toward extending EMVA 1288 to event-driven and photon-counting cameras.

References

1. A. Manakov, H. Herrmann, and B. Jähne, "Performance comparison of area-scan and event-based camera," *Tagungsband "Forum Bildverarbeitung 2024*, 2024.
2. A. Manakov, H. Herrmann, and B. Jähne, "Characterisation metrics for performance comparison of area-scan and event-based sensors," *tm - Technisches Messen*, 2025.
3. G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.
4. C. Posch and D. Matolin, "Sensitivity and uniformity of a 0.18µm CMOS temporal contrast pixel array," in *Circuits and Systems, ISCAS 2011, IEEE International Symposium*, 2011, pp. 1572–1575.
5. I. Cusini, D. Berretta, E. Conca, A. Incoronato, F. Madonini, A. A. Maurina, C. Nonne, S. Riccardo, and F. Villa, "Historical perspectives, state of art and research trends of single photon avalanche diodes and their applications (part 1: Single pixels)," *Frontiers in Physics*, vol. Volume 10 - 2022, 2022.
6. F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.
7. I. Cusini, D. Berretta, E. Conca, A. Incoronato, F. Madonini, A. A. Maurina, C. Nonne, S. Riccardo, and F. Villa, "Historical perspectives, state of art and research trends of SPAD arrays and their applications (part ii: SPAD arrays)," *Frontiers in Physics*, vol. Volume 10 - 2022, 2022.
8. A. Manakov and B. Jähne, "Characterization of event-based image sensors in extent of the EMVA 1288 standard," *Tagungsband "Forum Bildverarbeitung 2020*, 2020.
9. A. Manakov, H. Herrmann, and B. Jähne, "Characterization of event-based image sensors in extent of the EMVA 1288 standard," in *6th European Machine Vision Forum*, Wagingen, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:269583737
10. C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 db dynamic range asynchronous address-event pwm dynamic image sensor with lossless pixel-level video compression," in *IEEE Conference of Solid-State Circuits*, 2010, pp. 400–401.

Multipoints - Eine Übersicht

Multipoints - An overview

Tobias Haist, Nizar Rustom, Michael Tscherpel, Andreas Fuhrmann und Stephan Reichelt

Institut für Technische Optik, Universität Stuttgart
Pfaffenwaldring 9, 70569 Stuttgart

Zusammenfassung Die bildbasierte Messung zweidimensionaler Punktpositionen bildet die Grundlage vieler optischer Messverfahren und ist in der praktischen Messtechnik von zentraler Bedeutung. Üblicherweise werden heute dabei Bildsensoren zur genauen Bestimmung der Position eingesetzt. In der Praxis lässt sich oft mit einer einfachen Modifikation der Optik durch die sogenannte Multipointmethode die Genauigkeit wesentlich steigern. Die Multipointmethode lässt sich mit „analoge Replikation vor der Fotodetektion" zusammenfassen. Im Kern handelt es sich um einen vielseitigen Ansatz zur räumlichen Mittelwertbildung bei bildsensorbasierten Messungen, der die punktweise Messfähigkeit bewahrt. Rauschen (z.B. Photonenrauschen, thermisches Rauschen) und Pseudorauschen (z.B. Fehler aufgrund der Diskretisierung) werden reduziert. In diesem Beitrag wird die Grundidee motiviert, die Theorie und die bildbasierte Auswertung erläutert und es wird auf eine der vielen Anwendungen zur hochgenauen Positionsbestimmung in mechanischen Verfahreinheiten etwas näher eingegangen.

Schlüsselwörter Mechanische Verfahreinheiten, Punktdetektion, Bildverarbeitung, optische Messtechnik, Multipoints

Abstract Image-based measurement of two-dimensional point positions forms the basis of many optical measurement methods and is of central importance in practical metrology. Today, image sensors are typically used to determine the position with high accuracy. In practice, accuracy can often be significantly improved with a simple modification to the optics using the so-called multipoint method. The multipoint method can be summarized as

„analog replication prior to photodetection“. At its core, it is a versatile approach to spatial averaging in image-sensor-based measurements that preserves the point-by-point measurement capability. Noise (e.g., photon noise, thermal noise) and pseudo-noise (e.g., errors due to discretization) are reduced. This article explains the underlying concept, details the theory and image-based evaluation, and discusses one of the many applications for high-precision position determination in mechanical motion units in somewhat greater detail.

Keywords Mechanical stages, point detection, multipoint

1 Grundlagen und Anwendungen

In vielen optischen Messverfahren müssen intern im zweidimensionalen Bild Positionen gemessen werden. Beispiele sind klassische 2D- und 3D-Bildverarbeitungsaufgaben, bei denen die Geometrie z.B. von Werkstücken erfasst werden soll oder auch abgelenkte Laserspots, z.B. zur Messung von Neigungen.

Wenn dabei keine extremen Anforderungen an Geschwindigkeit und Latenz der Positionsbestimmung gestellt werden, dann werden heute meist bildsensorbasierte Methoden eingesetzt. Im Vergleich zu analogen Verfahren, wie z.B. 4-Quadrantendioden, erzielt man eine höhere relative Genauigkeit und die Flexibilität steigt. Nicht zuletzt können eben mehrere Positionen gleichzeitig von einem Bildsensor erfasst werden, so dass einfach auch z.B. relative Messungen realisierbar sind.

Klassischerweise wird bei entsprechenden bildsensorbasierten Messungen ein möglichst gutes optisches Signal (Elimination von Störlicht durch Filterung) gewonnen und dann z.B. mittels Schwerpunktsbildung die Position bestimmt. In der Praxis werden hier Genauigkeiten von 1/10-tel bis 1/100-tel eines Pixel erreicht. Bei den heute preisgünstig verfügbaren hohen Ortsbandbreiten (Pixel) von Bildsensoren ergibt sich so dann schnell eine recht hohe relative Genauigkeit.

Begrenzt wird die Genauigkeit (bei ausreichend guter Kalibrierung) durch restliche zufällige bzw. pseudo-zufällige Fehler, also vor allem Lichtrauschen, Bildsensorrauschen, Diskretisierungsfehlern.

Die Multipointmethode nutzt nun einen sehr einfachen Trick, um

diese Fehleranteile zu verringern: *Analoge Replikation vor der Fotodetektion!*

Konkret wird in den (Abbildungs-)Strahlengang ein optisches Element eingebracht, das die Information räumlich versetzt mehrfach kopiert [1]. So ergeben sich auf dem Bildsensor also an verschiedenen Stellen die gleichen Bilder und diese Bilder können dann getrennt ausgewertet werden. Zufällige und zueinander unkontrollierte Fehler werden dabei um den Faktor $\sqrt{N}$ reduziert, wenn wir mit N die Anzahl der Kopien bezeichnen.

Noch konkreter verdeutlicht Abb. 1 eine einfache optische Modifikation, die zum gewünschten Ergebnis führt: Ein computergeneriertes Hologramm führt (Beugung an den elementaren Gittern des Hologramms) zu N Lichtablenkungen und damit den N Kopien des Lichts.

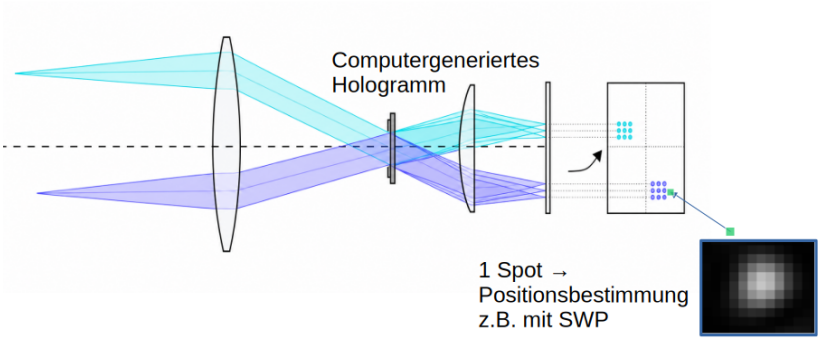


Abbildung 1: Grundidee der Multipointmethodik hier umgesetzt mit einem computergenerierten Hologramm und einem doppelt-telezentrischen Abbildungssystem.

Die Methodik wurde bereits erfolgreich in verschiedenen dafür geeigneten optischen Messverfahren eingesetzt, die im Folgenden kurz vorgestellt werden.

1.1 Deformationsmessungen an großen Strukturen

Die Deformationsmessung großer Strukturen und Bauwerken, wie z.B. Brücken und Hochhäusern ist zum einen im Sinne einer Überwachung, sogenanntes structural health monitoring, interessant, ist aber auch zur Steuerung adaptiver Strukturen wichtig.

Hartlieb, Guerra et al. haben in verschiedenen Arbeiten gezeigt, dass die Multipointmethode hier deutliche Vorteile gegenüber anderen technischen Ansätzen hat [2–5]. Insbesondere wenn eine kontinuierliche und zeitlich hochaufgelöste Überwachung bei geringen Kosten angestrebt wird ist eine kamerabasierte Lösung eigentlich konkurrenzlos. An die Struktur werden einzelne Lichtemitter, typischerweise LEDs, angebracht und mittels einer Kamera die LED Positionen ermittelt. Es lassen sich somit die relativen Verschiebungen von einzelnen Gebäudepunkten zueinander bestimmen.

Bei großen Gebäuden ergibt sich das Problem, dass mit einer einzelnen Kamera entweder ein nur stark begrenzter Bereich abgebildet werden kann, oder aber die Genauigkeit der Erfassung schlecht wird. Klassischerweise wird dann zu Netzwerken aus Kameras gewechselt [6], allerdings ergibt sich dann das Problem, dass minimale Änderung der Pose von Kameras zueinander zu Messfehlern führt.

Die Multipointmethode vermeidet diese Nachteile und es wird genutzt, dass auch mit einer Kamera eine hohe relative Genauigkeit erzielt werden kann. Die Methodik wurde erfolgreich zur Deformationsmessung an Gebäuden und Brücken demonstriert. In [7] wurden dabei für eine Bahnbrücke bei Hamburg Deformationen mit Genauigkeiten kleiner 100 μm erzielt.

1.2 Vibrometrie

Aufgrund der hohen relativen Genauigkeit der Multipointmethode lassen sich bei kleinen Objekten sehr hohe Empfindlichkeiten und Genauigkeiten erzielen. Damit wird es möglich, Vibrationsmessungen durchzuführen, die klassischerweise mit kohärenten Verfahren, meist Laser-Doppler-Interferometrie, gelöst werden [8]. Da wieder lediglich ein Bildsensor und eine einfache Optik benötigt wird, kann wieder eine extrem kostengünstige Messung realisiert werden. Selbst mit sehr preiswerten Bildsensoren lassen sich in Regions-of-Interest Messraten im Be-

reich von kHz erzielen. Für höherfrequente Schwingungen müssen entsprechend teurere Hochgeschwindigkeitskameras eingesetzt werden.

Als Einschränkung muss genannt werden, dass wieder auf dem schwingenden Objekt aktive Emitter (sinnvollerweise meist Mikro-LEDs) angebracht werden. Dies ist ein Nachteil gegenüber laserbasierten Verfahren. Dafür kann die Multipointmethode wieder den Vorteil ausspielen, dass mehrere Punkte gleichzeitig gemessen werden können und daher eben Relativmessungen und damit Phasenmessungen durchgeführt werden können; etwas, das mit klassischen Vibrometern nur bei stationären Schwingungen und nicht bei transienten Schwingungen einfach realisierbar ist.

1.3 Shack-Hartmann Sensorik

Ein ganz anderes Beispiel zeigt gut, dass die Multipointmethodik in ganz unterschiedlichen optischen Messsystemen vorteilhaft eingesetzt werden kann. Hartlieb et al. haben demonstriert, dass der klassische Shack-Hartmann Sensor (SHS) einfach durch die Multipointmethode verbessert werden kann.

Die einzelne Apertur des SHS, die normalerweise auf dem Bildsensor einen Spot generiert, wird ersetzt durch ein computergeneriertes Hologramm, das ein komplexeres Punktemuster generiert [9].

Dadurch kann man — im Sinne der Multipointmethodik — wieder die Genauigkeit durch Mittelung erhöhen. Außerdem wird die Dynamik des Sensors größer, weil einzelne Muster nun über einen größeren Bereich auf dem Bildsensor eindeutig auffindbar sind.

1.4 Triangulation und ausgedehnte Bilder

Die Steigerung der Genauigkeit durch Vervielfachung kann natürlich auch für triangulierende Systeme genutzt werden. Man denke an eine einfache Lasertriangulation: Hier ist die Genauigkeit wesentlich (auch) von der Positionsbestimmung in der Bildebene abhängig. Statt einen Punkt auszuwerten werden wieder viele Punkte ausgewertet und die Ergebnisse werden gemittelt. Faulhaber und Haist et al. haben hiermit z.B. die Robustheit von kamerabasierter Atmungsdetektion gesteigert [10, 11].

Bisher sind wir immer von einzelnen isolierten Emittern (z.B. LEDs) ausgegangen, die auf den Bildsensor (repliziert) abgebildet werden. Natürlich lässt sich die Methode aber auch auf ausgedehnte Strukturen anstatt punktartige Objekte erweitern. Allerdings muss dann für die meisten Szenen eine Zwischenabbildung mit Zwischenbildebene bei der Abbildung auf den Bildsensor eingebaut werden, so dass nur die Teile (repliziert) abgebildet werden, die informationstragend sind [12, 13].

1.5 Koordinatenmessmaschinen und Bearbeitungsmaschinen

Auch hochgenaue Mess- und Bearbeitungsmaschinen können von der Technik profitieren. Konventionelle Maschinen nutzen zur Erzielung hoher Genauigkeiten vor allem Glasmaßstäbe in den Achsen. Sobald aber statische oder dynamische Lasten zu leichten Deformationen führen, wird es schnell schwierig, Genauigkeiten im einstelligen Mikrometerbereich zu erzielen; ganz besonders bei vielen Achsen.

Die Multipointmethodik umgeht diese Probleme indem eben das gemessen werden kann, was wirklich interessiert, nämlich die Position (oder auch Orientierung) nicht in einem globalen Koordinatensystem der Maschine, sondern die relative Position zwischen Objekt und Bearbeitungs- oder Messkopf. Es werden entsprechend Emitter an der Werkzeugaufnahme und der Objektaufnahme angebracht [14, 15]. Hartlieb et al. haben auch eine Erweiterung auf 3D Messungen gezeigt [16].

Im Folgenden werden wir in diesem Beitrag etwas ausführlicher auf eine aktuelle Erweiterung bzw. Abwandlung dieser Anwendung eingehen, nämlich die hochgenaue Regelung von mechanischen Verfahreinheiten.

2 Holostage

Leider steigen Aufwand und Kosten mechanischer Verfahreinheiten stark mit der gewünschten Genauigkeit an. Ganz besonders gilt das für mehrachsige Systeme. Üblicherweise müssen ab einer gewissen Genauigkeit Maßstäbe (z.B. aus Glas) in den Achsen der Maschine verwendet werden. Bei mehreren Achsen, Kraftbelastungen und dynamischen

Anforderungen ergeben sich so aber oft schnell massive Probleme hinsichtlich der Genauigkeit.

Dies zeigt sich auch an bisherigen kommerziellen Verfahreinheiten, bei denen oft zwar eine sehr gute Auflösung (z.B. sub-Mikrometer) aber eben eine geringe Genauigkeit (einige zehn Mikrometer) erreicht werden.

2.1 Grundidee

Die Holostage vermeidet diese Probleme, indem eine bildbasierte, per Multipoint vervielfachte hochgenaue Positionsbestimmung der Verfahreinheiten verwendet wird. Das grundsätzliche Prinzip ist in Abb. 2 dargestellt. Dabei wird jeweils eine Mikro-LED (Abb. 3) eines LED-Arrays vervielfacht auf den Bildsensor abgebildet.

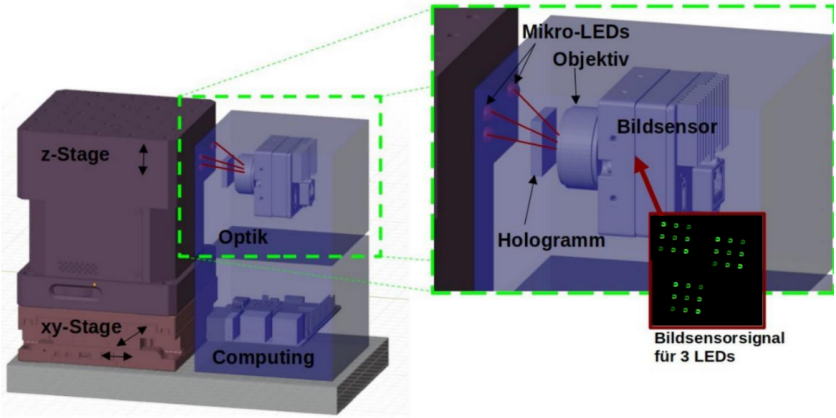


Abbildung 2: Grundidee Holostage: Eine mechanische Stage geringer Genauigkeit wird mit dem Multipointverfahren zur genauen Verfahreinheit.

Idealerweise soll eine Genauigkeit von 1 µm erreicht werden, so dass die Positionsbestimmung im Bild auf 0.1 Pixel genau erfolgen muss. Das Abbildungssystem nutzt objektseitige Telezentrie. Hier wird als

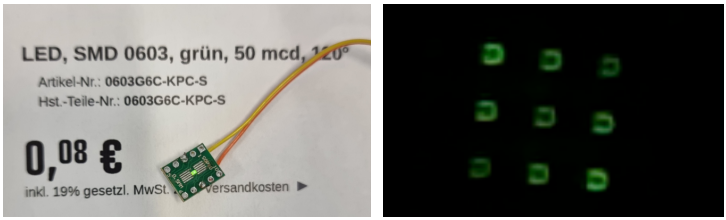
Abstand zwischen erster Linse (Achromat, 35 mm) und der Eintrittspupille des nachfolgenden Pi Cam Moduls (4,7 mm Brennweite) die Brennweite (35 mm) gewählt. Dadurch laufen die Strahlenbündel im Objektraum nahezu parallel.

Beim aktuellen Prototyp erfolgt eine optische Abbildung mit einem Abbildungsmaßstab von -0.13. Das bedeutet, dass eine Verschiebung von 0.1 Pixel beim eingesetzten Bildsensor (Raspberry Pi Cam modul V3, Sony IMX708, 4608 x 2592 Pixel, Pixelgröße 1.4 µm) einer Verschiebung von ca. 1 µm entspricht.

Als computergeneriertes Hologramm wird ein binäres Phasenelement verwendet. Dies hat den Vorteil, dass die positiven und negativen Beugungsordnungen gleich hell sind und so keine Asymmetrie in der Positionsauswertung auftreten kann. Außerdem ist die Herstellung einfach und preiswert.

2.2 Deep learning und Korrelation

Für die Positionsbestimmung können sehr verschiedene Varianten zum Einsatz kommen. Zu beachten ist dabei, dass das Bild einer LED zum einen durch Aberrationen und Beugung modifiziert ist, vor allem aber Mikro-LEDs je nach Kapselung durchaus inhomogene Leuchtdichteverteilungen haben. Ein typisches Bild auf dem Bildsensor sieht damit wie in Abb. 3 dargestellt aus.



(a) Mikro-LED **(b)** Ausschnitt Bildsensor

Abbildung 3: Vervielfachtes Punktbild der Mikro-LED auf dem Bildsensor.

Die Positionsbestimmung erfolgt dabei zunächst separat für jede

Bildkopie; jede Bildkopie liefert also eine Position (x, y). Die finale Position ergibt sich dann durch Mittelwertbildung. Wichtig ist dabei, dass sich potentiell systematische Fehler möglichst gegenseitig ausgleichen.

Die Positionsbestimmung erfolgt zweigeteilt. Zunächst wird mit verschiedenen Verfahren eine grobe, pixelgenaue Position ermittelt. Dann wird eine Region-of-Interest um die Position gebildet und jeweils der Schwerpunkt bestimmt. Dabei sollten Werte unterhalb einer Schwelle (z.B. Mittelwert im Spotbereich) nicht beachtet bzw. auf Null gesetzt werden.

Eine konturbasierte, eine Blob-basierte und eine korrelationsbasierte Variante wurden untersucht. Es zeigte sich, dass mit der korrelationsbasierten Methodik die besten Ergebnisse, im Genauigkeitsbereich von 0.04 Pixel (Standardabweichung) erzielt werden können. Dies entspricht bei der eingesetzten Optik und dem Bildsensor mit 1.4 µm großen Pixeln einer Genauigkeit von ca. 0.4 µm im Objektraum.

Zur Genauigkeitssteigerung ist es sinnvoll, vor der Auswertung eine Interpolation (4-fach Lanczos) durchzuführen. Dadurch wird zwar die Rechenzeit erhöht, aber eben auch die Genauigkeit verbessert.

Auf einem Raspberry Pi 5 ergeben sich bei 9 Bildern Rechenzeiten (unoptimiertes Python) im Bereich 15 ms.

Auf lernenden Verfahren beruhende Auswertungen können eine interessante Alternative zur klassischen Bildverarbeitung darstellen. In ersten Versuchen wurden verschiedene Ansätze getestet. In Verbindung mit Korrelation konnte — mit allerdings einem verringerten Verfahrbereich von 70 µm— eine im Vergleich zur Schwerpunktsauswertung um den Faktor 4 höhere Genauigkeit mit der in Abb. 4 dargestellten Struktur erzielt werden. Im Prinzip können damit auch gleich systematische Fehler erfasst bzw. ausgeglichen werden. Die Rechenzeit war um ca. 40% geringer als für die Auswertung mittels Korrelation und Schwerpunkt.

3 Zusammenfassung

Das Multipointverfahren ist eine in der optischen Messtechnik breit anwendbare Grundidee. Immer dann, wenn Positionen auf zweidimensionalen Bildsensoren hochgenau ermittelt werden müssen, kann — vorausgesetzt genügend Licht ist vorhanden — eine Replikati-

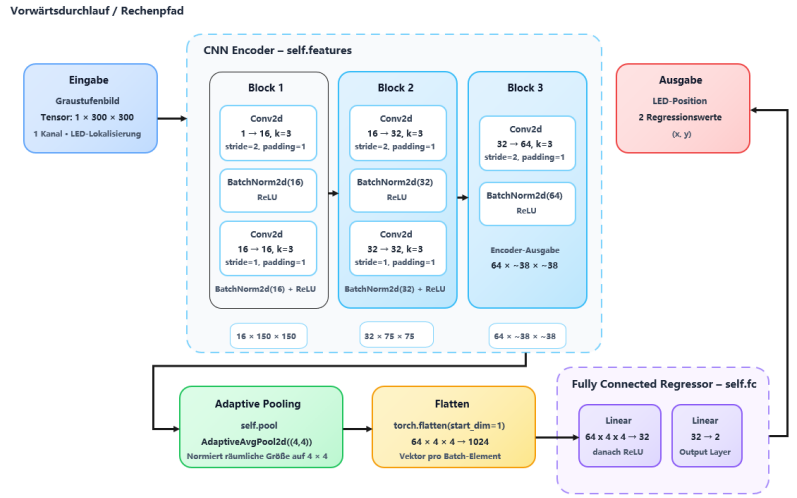


Abbildung 4: Deep Learning Ansatz zur Auswertung.

on der optischen Information vor der eigentlichen Bildaufnahme genutzt werden, um zufällige (z.B. Lichtrauschen, thermisches Rauschen) und pseudozufällige (z.B. Diskretisierungsrauschen durch Pixelierung) Fehleranteile deutlich zu reduzieren. Dabei muss natürlich sichergestellt werden, dass sich die wiederholten Kopien auf dem Bildsensor nicht überlappen.

Unter anderem ist das Verfahren nützlich, um die Genauigkeit von mechanischen Verfahreinheiten deutlich zu steigern. Mit einfachen Mitteln kommt man so nach Kalibrierung in den Bereich unter 1 µm.

Acknowledgements

Wir danken dem Bundesministerium für Wirtschaft und Energie (BMWE) für die finanzielle Unterstützung dieser Arbeit unter dem Förderkennzeichen KK 5051809.

Literatur

1. T. Haist, S. Dong, T. Arnold, M. Gronle, and W. Osten, "Multi-image position detection," *Optics express*, vol. 22, no. 12, pp. 14 450–14 463, 2014.
2. S. Hartlieb, M. Tscherpel, F. Guerra, T. Haist, W. Osten, M. Ringkowski, and O. Sawodny, "Highly accurate imaging based position measurement using holographic point replication," *Measurement*, vol. 172, p. 108852, 2021.
3. F. Guerra, T. Haist, A. Warsewa, S. Hartlieb, W. Osten, and C. Tarín, "Precise building deformation measurement using holographic multipoint replication," *Applied Optics*, vol. 59, no. 9, pp. 2746–2753, 2020.
4. F. Guerra, S. Hartlieb, A. Warsewa, T. Haist, W. Osten, and O. Sawodny, "Deformation measurement of large buildings by holographical point replication," in *Optics for Arts, Architecture, and Archaeology VII*, vol. 11058. SPIE, 2019, pp. 32–38.
5. F. Guerra, A. Warsewa, S. Hartlieb, T. Haist, W. Osten, and O. Sawodny, "Holographic point replication as a sensor-enhancing technique for adaptive building control," in *Photonic Instrumentation Engineering VII*, vol. 11287. SPIE, 2020, pp. 209–217.
6. J.-W. Park, J.-J. Lee, H.-J. Jung, and H. Myung, "Vision-based displacement measurement method for high-rise building structures using partitioning approach," *Ndt & E International*, vol. 43, no. 7, pp. 642–647, 2010.
7. S. Hartlieb, A. Zeller, T. Haist, A. Reichardt, C. Tarín Sauer, and S. Reichelt, "Advanced imaging-based metrology for precise deformation monitoring: railway bridge case study," *Structural Control and Health Monitoring*, vol. 2025, no. 1, p. 5603393, 2025.
8. S. Hartlieb, M. Boguslawski, T. Haist, and S. Reichelt, "Holographical image based vibrometry with monochromatic and event based cameras," in *Optics and Photonics for Advanced Dimensional Metrology II*, vol. 12137. SPIE, 2022, p. 1213702.
9. S. Hartlieb, Z. Wang, A. Rüdinger, T. Haist, and S. Reichelt, "Large dynamic range shack–hartmann wavefront sensor based on holographic multipoint generation and pattern correlation," *Optical Engineering*, vol. 63, no. 2, pp. 024 107–024 107, 2024.
10. A. Faulhaber, M. Gronle, S. Haberl, T. Buchholz, T. Haist, and W. Osten, "Dynamically scanned spot projections with digital holograms for reduced measurement uncertainty in laser triangulation systems," in *AOPC 2019: Optical Sensing and Imaging Technology*, vol. 11338. SPIE, 2019, pp. 856–868.

11. T. Haist, C. Reichert, F. Würtenberger, L. Lachenmaier, and A. Faulhaber, "Kamerabasierte Erfassung von Vitalparametern," *tm-Technisches Messen*, vol. 86, no. 7-8, pp. 354–361, 2019.
12. S. Hartlieb, M. Ringkowski, T. Haist, O. Sawodny, and W. Osten, "Multi-positional image-based vibration measurement by holographic image replication," *Light: Advanced Manufacturing*, vol. 2, no. 4, pp. 425–433, 2021.
13. V. Aslani, F. Guerra, A. Steinitz, P. Wilhelm, and T. Haist, "Averaging approaches for highly accurate image-based edge localization," *Optics Continuum*, vol. 1, no. 4, pp. 834–845, 2022.
14. S. Hartlieb, M. Tscherpel, F. Guerra, T. Haist, W. Osten, M. Ringkowski, and O. Sawodny, "Hochgenaue Kalibrierung eines holografischen Multi-Punkt-Positionsmesssystems," *tm-Technisches Messen*, vol. 87, no. 7-8, pp. 504–513, 2020.
15. S. Hartlieb, C. Erol, M. Tscherpel, T. Haist, F. Guerra, W. Osten, M. Ringkowski, and O. Sawodny, "Accurate 3d coordinate measurement using holographic multipoint technique," in *Optics and Photonics for Advanced Dimensional Metrology*, vol. 11352. SPIE, 2020, p. 1135203.
16. S. Hartlieb, C. Schober, T. Haist, and S. Reichelt, "Accurate single image depth detection using multiple rotating point spread functions," *Optics Express*, vol. 30, no. 13, pp. 23 035–23 049, 2022.

Comparative analysis of integral transform-based single- and few-shot fringe image processing algorithms

Ömer Atmaca, Wala Hasnaoui, Mina Ebrahimnejad, and Alois Herkommer

University of Stuttgart, Institute of Applied Optics (ITO), Pfaffenwaldring 9, 70569 Stuttgart

Abstract This work provides a comparison between integral transform-based fringe image processing algorithms for fringe phase reconstruction. This analysis includes an automatic Radon transform-based Fourier transform profilometry algorithm and compares it against the dual-tree complex wavelet transform and the Riesz transform profilometry. The algorithms are benchmarked on fringe images with different levels of noise and artifacts. They are furthermore evaluated on practicality, ease of implementation, time complexity, and the minimum amount of images needed.

Keywords Fringe projection, Fourier transform, phase, profilometry, Radon transform, Riesz transfrom, wavelet transform

1 Introduction and description of algorithms

Fringe projection offers valuable insights into object properties, such as surface profiles or optical scattering and absorption characteristics. Processing many phase-shifted fringe images improves the reconstruction quality. However, many 3D applications in industrial and biomedical contexts require video-frame processing capabilities. A sizable training dataset for robust data-driven reconstruction based on a single image is often not available. Hence, single- and few-shot fringe image processing methods based on integral transforms, like Fourier, Riesz, or wavelet transform profilometry, are frequently in use. There, a fringe

image $I(\mathbf{r}) = I_{\mathrm{DC}}(\mathbf{r}) + I_{\mathrm{AC}}(\mathbf{r}) = a(\mathbf{r}) + b(\mathbf{r})\cos(2\pi\mathbf{f_0}\cdot\mathbf{r} + \varphi(\mathbf{r}))$ of size $M, N \in \mathbb{N}_0$ (Fig. 1 (a)) with its so-called *DC/bias* and *AC components* contains the carrier frequency $\mathbf{f_0} = (f_{0,x}, f_{0,y})$ and phase φ. I is processed and/or demodulated at every pixel $\mathbf{r} = (x, y)$ in the transform domain for their desired quantity like the fringe phase, which is directly linked to the surface height profile of an object. However, these methods usually have the disadvantage of being less robust or relying on manually tuned static parameters, which can substantially impact the reconstruction quality. This work therefore compares an automatic Radon transform (RaT)-based Fourier transform (FT) profilometry algorithm as a step towards parameter-free fringe image evaluation, against the dual-tree complex wavelet transform (DTCWT), and variants of Riesz transform (RiT) profilometry. The algorithms are benchmarked on different fringe images with different levels of noise and artifacts and are evaluated on different metrics like practicality, ease of implementation, time complexity, and the minimum amount of images needed.

1.1 Radon transform-based Fourier transform profilometry (RTFTP)

Fourier transform profilometry (FTP) by Takeda *et al.* [1] demodulates the fringe image I in the spatial frequency domain (SFD) (Fig. 1 (b)): $I(\mathbf{f}) = A(\mathbf{f}) + C(\mathbf{f} - \mathbf{f_0}) + C^*(\mathbf{f} + \mathbf{f_0})$ by band-pass filtering either one of the carrier frequency components (C or C^*), separating it from the zero frequency spectrum A, and shifting it to the origin of the SFD.

In the RTFTP, this filtering is performed with a high-pass, then a low-pass, and then a band-pass filter $H, L, B \in \{0,1\}^{M\times N}$ such that $C = B \odot L \odot H \odot I$. The parameters for H, L, and B can be determined automatically from the sinogram R of the power spectral density of $I(\mathbf{r})$. The forward RaT of $|I(\mathbf{f})|^2$ results in three sinusoidal bands $R_{|I|^2}(f,\theta) = R_{|A(\mathbf{f})|^2+|C(\mathbf{f}-\mathbf{f_0})|^2+|C^*(\mathbf{f}+\mathbf{f_0})|^2}(f,\theta)$, with the projection angle θ and $f = |\mathbf{f}|$.

The first feature of interest (FOI) is the orientation of the fringe. The angle of the fringe direction θ_1 is where A, C, and C^* are projected onto each other in $R_{|I|^2}$ and is therefore the angle where $R_{|I|^2}$ becomes the sparsest. The sparsity is computed as a combined probability (density) function (PF) using four sparsity/entropy measures evaluated on

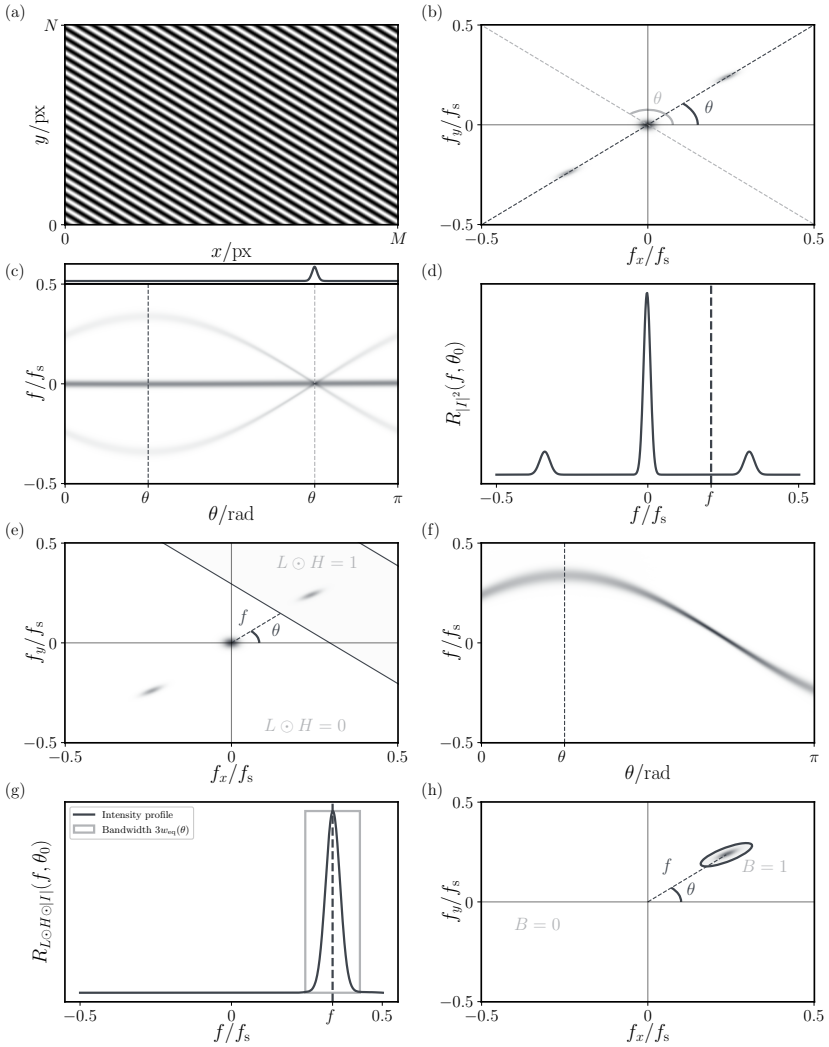


Figure 1: Principle of the RTFTP. (a) Fringe image I, (b) FT of (a) with projection lines at θ_0 and θ_1, (c) RaT of the squared magnitude of (b) with the normalized probability of at every θ on top, (d) Intensity profile of (c) at θ_0 with f_c, (e) High-pass H and low-pass filter L, (f) RaT of (e), (g) Intensity profile of (f) at θ_0 with f_0 and the bandwidth, (h) Band-pass filter B.

$R_{|I|^2}(f,\theta)$ over all f at every θ: the Hoyer sparsity, the pq-means for $p=1$ and $q=2$, the Gini coefficient, and the negative of the Shannon entropy. All four curves are min-max normalized and multiplied with each other, to then divide the result by its area for the PF. Then, θ_1 is the angle with the highest probability of the combined PF and the direction of the fringe wave number is $\theta_0 = (\theta_1 - \frac{\pi}{2}) \bmod \pi$ (Fig. 1 (c)).

$R_{|I|^2}(f,\theta_0)$ yields a trimodal distribution. To robustly find any relevant FOI on this distribution, the concept of topographic peak salience is introduced: If $\mathcal{C}(g_\mathcal{C}(x)) = \{x_{\mathcal{P},i}\}$ returns the positions of topographic peaks $x_{\mathcal{P},i}$ of a candidate function $g_\mathcal{C}$ as the candidate points for the FOI and if $P = \mathcal{P}(g(x), \{x_{\mathcal{P},i}\})$ returns the prominences P of these candidate points on the function g, then the peak salience is $\mathcal{S}(g(x), g_\mathcal{C}(x)) = \mathcal{P}(\mathcal{P}(g(x), \mathcal{C}(g_\mathcal{C}(x))), \mathcal{C}(\mathcal{P}(g(x), \mathcal{C}(g_\mathcal{C}(x)))))$.

The second FOI is the cutoff frequency f_c for H and L, which is the most salient valley of $R_{|I|^2}(f,\theta_0)$ between the major mode ($f=0$) and the minor mode ($f \approx f_0$) (Fig. 1 (d)). The candidate function for the most salient peak of the minor mode is $\{g_{\mathcal{C},1}, g_{\mathcal{C},2}\} = \left\{R_{|I|^2}(f,\theta_0), \max\left\{-\mathrm{d}^2 R_{|I|^2}(f,\theta_0)/\mathrm{d}f^2, 0\right\}\right\}$ with $\mathcal{C}(\{g_{\mathcal{C},1}, g_{\mathcal{C},2}\}) = \mathcal{C}(g_{\mathcal{C},1}) \cup \mathcal{C}(g_{\mathcal{C},2})$ for $f > 0$. $g_{\mathcal{C},2}$ is computed with a Savitzky-Golay filter with a globally optimized filter window and polynomial order [2]. Therefore, $f_\mathrm{c} = \operatorname{argmax}\left(\mathcal{S}\left(-R_{|I|^2}(f,\theta_0), -R_{|I|^2}(f,\theta_0)\right)\right)$ for $f \in (0, f_0)$. If no salient local minimum is found, $f_\mathrm{c} \approx \frac{f_0}{2}$.

(f_c, θ_0) are polar coordinates in the SFD to suppress A and C^* with $H = \mathbb{1}_{d \geq \frac{4}{3} f_\mathrm{c}} + \frac{1}{2}\left(1 - \cos\left[\frac{3\pi(f-f_\mathrm{c})}{f_\mathrm{c}}\right]\right) \mathbb{1}_{f_\mathrm{c} \leq d < \frac{4}{3} f_\mathrm{c}}$ and $d = f_y \sin(\theta_0) + f_x \cos(\theta_0)$, where $\mathbb{1}_X$ is the indicator function on the set X (Fig. 1 (e)). The shape of the cosine fading for smooth filtering can be derived by considering the bandwidth limitations in FTP [1] and $f_0 \approx 2 f_\mathrm{c}$. L is defined similarly to H for $d \in \left(\frac{8}{3} f_\mathrm{c}, 3 f_\mathrm{c}\right]$.

A second RaT of $L \odot H \odot |I| \approx |C|$ results in a single sinusoidal band with $R_{L \odot H \odot |I|}(f,\theta) \approx R_{|C|}(f,\theta)$ (Fig. 1 (f)), which enables the band-pass filter design (as an ellipse for example) with $B = \mathbb{1}_{f_x'^2/f_\mathrm{maj}^2 + f_y'^2/f_\mathrm{min}^2 \leq 1}$ and $f_x' = (f_x - f_{0,x})\cos(\theta_\mathrm{maj}) + (f_y - f_{0,y})\sin(\theta_\mathrm{maj})$ and $f_y' = (f_x - f_{0,x})\sin(\theta_\mathrm{maj}) - (f_y - f_{0,y})\cos(\theta_\mathrm{maj})$. The remaining FOIs are therefore the filter parameters for B. $f_0 = |\mathbf{f_0}| = \operatorname{argmax}_{f_\mathrm{c} < f < 3 f_\mathrm{c}}\left(\mathcal{S}\left(R_{H \odot |I|^2}(f,\theta_0), R_{H \odot |I|^2}(f,\theta_0)\right)\right)$

or $f_0 = \operatorname{argmax}_f \left(R_{H \odot |I|^2}(f, \theta_0) \right)$ if no salient peak is found (Fig. 1 (g)). The equivalent rectangle width at every angle $w_{\mathrm{eq}}(\theta) = \int R_{L \odot H \odot |I|}(f, \theta) \mathrm{d}f / R_{L \odot H \odot |I|}(f_0, \theta)$ sets the bandwidth (Fig. 1 (h)). The principle semiaxes of the elliptical filtering window are therefore $f_{\mathrm{min}} = \frac{3}{2} w_{\mathrm{eq}}(\theta_{\mathrm{min}}) = \frac{3}{2} \min \left(w_{\mathrm{eq}}(\theta) \right)$ and $f_{\mathrm{maj}} = \frac{3}{2} w_{\mathrm{eq}}(\theta_{\mathrm{maj}}) = \frac{3}{2} w_{\mathrm{eq}} \left((\theta_{\mathrm{min}} - \frac{\pi}{2}) \bmod \pi \right)$. Anisotropic Gaussian blurring on B mitigates the Gibbs effect in the inverse FT of C to c. The filter window size in f_x and f_y are $3 w_{\mathrm{eq}}(0)$ and $3 w_{\mathrm{eq}} \left(\frac{\pi}{2} \right)$ (rounded to the next nearest odd integer) with standard deviations $\sigma_{C,\theta} = w_{\mathrm{eq}} (\theta) / \sqrt{2\pi}$.

The 2π-wrapped phase φ is then $\varphi(\mathbf{r}) = \arctan \left(\frac{\mathrm{Im}\{c(\mathbf{r})\}}{\mathrm{Re}\{c(\mathbf{r})\}} \right)$.

1.2 Dual-tree complex wavelet transform (DTℂWT) for bias removal

The bias removal using the DTℂWT by Ng *et al.* is especially effective in cases of fringe images with strong artifacts such as overexposed pixels (Fig. 2) [3, 4]. With additional noise $n \sim \mathcal{N}(0, \sigma_n^2)$ to $I(\mathbf{r})$, the transformed signal is $I^l_{\mathcal{W}} = a^l_{\mathcal{W}} + f^l_{\mathcal{W}} + n^l_{\mathcal{W}}$, where $\mathcal{W}$ describes the DTCWT operation. The fringe signal $f^l_{\mathcal{W}}$ concentrates in sub-band levels $l \in [l_{\mathrm{min}}, l_{\mathrm{max}}]$ with $l_{\mathrm{min}} = \left\lfloor \log_2 \left(\frac{f_{\mathrm{s}}}{f_0 + \frac{1}{2\pi} \left(\frac{\partial \varphi}{\partial x} \right)_{\mathrm{max}}} \right) \right\rfloor$ and $l_{\mathrm{max}} = \left\lceil \log_2 \left(\frac{f_{\mathrm{s}}}{2 \left(f_0 + \frac{1}{2\pi} \left(\frac{\partial \varphi}{\partial x} \right)_{\mathrm{min}} \right)} \right) \right\rceil$, where f_{s} is the sampling frequency [3]. For $l \notin [l_{\mathrm{min}}, l_{\mathrm{max}}]$, $I^l_{\mathcal{W}}$ is set to zero [4]. For $l_{\mathrm{min}} \leq l \leq l_{\mathrm{max}}$, the bias-free fringe signal $\hat{I}$ can be estimated using a soft-thresholding scheme of the form

$$\begin{aligned} \hat{I}^l_{\mathcal{W}} \approx \operatorname{med} \Bigg\{ \Bigg| I^l_{\mathcal{W}} - \max \Bigg\{ \sqrt{\max \left\{ \left| I^l_{\mathcal{W}} \right|^2 - \operatorname{med} \left\{ \left| I^l_{\mathcal{W}} \right| \right\}^2 + 2\sigma_n^2, 0 \right\}} \\ - \max \left\{ \sqrt{\frac{\pi}{2}} + 3 \sqrt{2 - \frac{\pi}{2}}, 6 \sqrt{\operatorname{med} \left\{ \left| I^l_{\mathcal{W}} \right| \right\}^2 - 2\sigma_n^2} \right\} \sigma_n, 0 \Bigg\} \\ \exp \left(\mathrm{j} \arg \left(I^l_{\mathcal{W}} \right) \right) \Bigg| \Bigg\} \exp \left(\mathrm{j} \arg \left(I^l_{\mathcal{W}} \right) \right), \end{aligned} \tag{1}$$

with $\mathrm{j}^2 = -1$. The median filter removes outliers in the signal and suppresses the bias terms. The remaining signal then only consists of

fringe and noise so that the rest becomes a simple denoising problem solved with a bilateral filter in the transform domain. The phase of the $\hat{I}$ can be calculated via RiT [5] (see the following section).

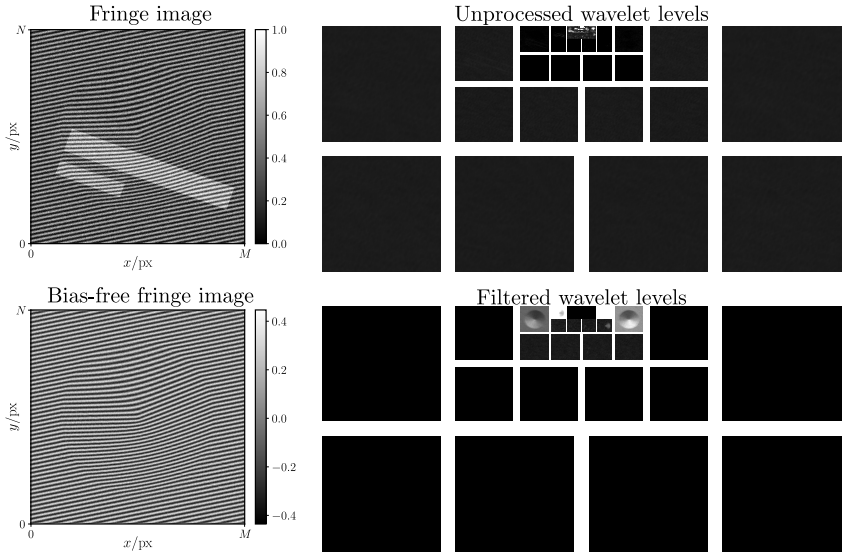


Figure 2: DTCWT bias removal. $f^l_{\mathcal{W}}$ concentrates in sub-band levels $l_{\text{min}} = 3$ & $l_{\text{max}} = 4$.

1.3 Riesz transform profilometry (RTP)

The RiT is the multidimensional generalization of the 1D Hilbert transform. Similarly, the generalization of the analytic signal (the monogenic signal) gives rise to the local amplitude and phase of the fringe image. This, however, requires I to be free of any high-frequency bias, which can be achieved for example with the DTCWT or by subtracting a second fringe image taken with $f_0 = 0$ (two-shot RTP, 2SRTP). The RiT on the $\hat{I}$ is applied via the two orthogonal kernels $r_v = \frac{v}{2\pi|\mathbf{r}|^3}$ with $v \in \{x, y\}$. The exact calculation of φ depends on the type of fringe pattern. For oriented fringes, it is for example possible to determine the phase from the structure tensor built

with the Riesz coefficients: $\varphi = \arctan\left(\frac{r_x * \hat{I}\cos(\theta_0) + r_y * \hat{I}\sin(\theta_0)}{\hat{I}}\right)$ with $\theta_0 = \frac{1}{2}\arctan\left(\frac{2(r_x * \hat{I})(r_y * \hat{I})}{(r_x * \hat{I})^2 - (r_y * \hat{I})^2}\right)$. It is also possible (even without orientation) to perform digital four-step phase-shifting (D4-PS) from a single pattern [6] and to use $\varphi = \arctan\left(\frac{I_{3\pi/2} - I_{\pi/2}}{I_0 - I_\pi}\right)$ with

$$\begin{aligned} I_0 &= \hat{I},\ I_{\frac{\pi}{2}} = \operatorname{sgn}(r_v * I_0)\sqrt{(r_x * I_0)^2 + (r_y * I_0)^2}, \\ I_\pi &= r_x * r_x * I_0 + r_y * r_y * I_0, \\ I_{\frac{3\pi}{2}} &= \operatorname{sgn}(r_v * r_v * r_v * I_0)\sqrt{(r_x * r_x * r_x * I_0)^2 + (r_y * r_y * r_y * I_0)^2}. \end{aligned} \tag{2}$$

2 Quantitative comparison

The following variants of the abovementioned algorithms were used for comparison: three phase-shifted profilometry (3PSP) as the standard, the RTFTP, the DTCWT with RTP for oriented fringes, the 2SRTP, and the D4-PS. The fringes were virtually projected onto two surfaces, namely a cone and the peaks function known from MATLAB (Fig. 3). Similarly to Ng *et al.* [3,4], different forms of artifacts were introduced to the fringe before performing phase reconstruction for sensitivity assessment: bias in the form of rectangles with different intensities and a patch of overexposed pixels to mimic specular reflections as well as Gaussian noise (see Tab. 1). Intensity values $I \notin [0,1]$ were clipped. The phase of a reference plane was subtracted modulo-2π from φ. Phase unwrapping was performed with PUMA [7]. Any residual planar phase slope and offset to the first image pixel were corrected with a least-squares fit to the four corners of the phase image.

The residual phase errors in Fig. 4 and in Tab. 1 show that all methods are capable of faithfully reconstructing phase of both objects. The reconstruction quality decreases with increasing levels of noise and artifacts. The artifacts of the form of bias and oversaturated patches have a substantially larger impact compared to the noise. The DTCWT has the best reconstruction quality among all methods, being only second to the 3PSP. The RTFTP behaves robustly even against strong artifacts and only substantially degrades at very high noise levels. Both

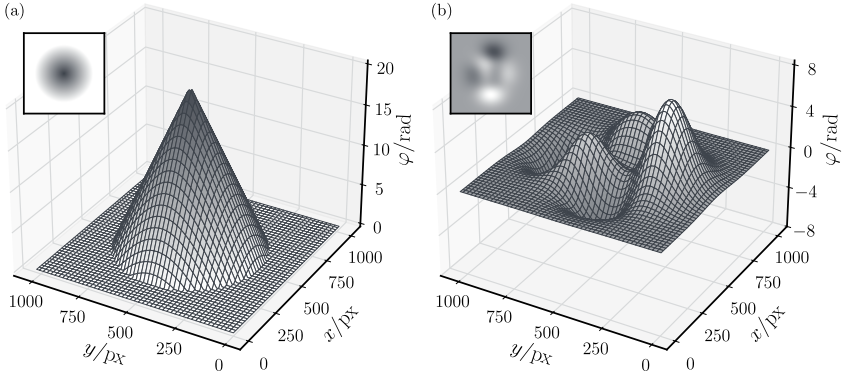


Figure 3: Simulated surfaces for fringe projection profilometry together with their phase maps. (a) Cone, (b) Peaks.

RTP-based methods exhibit good overall reconstruction quality while maintaining the spatial resolution, but suddenly drop once dominant high-frequency components in the form of sharp edges and noise are introduced.

Table 1: Root mean squared error (RMSE) values of the phase reconstructions in Fig. 4 with different forms of artifacts: rectangular bias terms, Gaussian noise σ_n, differently sized patches of overexposed pixels. The phase range magnitude of the cone object is 20 rad (see Fig. 3 (a)) and the phase range magnitude of the peaks object is 14.6572 rad (see Fig. 3 (b)).

Row	Object	Artifacts			Relative RMSE/%				
		Bias	σ_n	Patch	3PSP	RTFTP	DTCWT	2SRTP	D4-PS
1	Cone	✗	0	0	0.0	0.5112	0.1515	0.0166	0.069
2	Cone	✓	0.05	20	0.2497	0.564	0.3917	0.6279	1.4527
3	Cone	✓	0.1	20	0.3163	0.5935	0.441	0.6855	1.8991
4	Cone	✓	0.2	20	0.7233	0.6026	0.5701	1.3837	3.2951
5	Peaks	✗	0	0	0.0	0.724	0.2937	0.0558	0.4004
6	Peaks	✗	0	100	1.6767	1.118	0.8434	2.2847	2.3438
7	Peaks	✓	0	0	0.0014	0.841	0.3042	0.7639	1.5146
8	Peaks	✗	0.3	0	1.4538	2.4839	0.5718	2.1453	4.5736

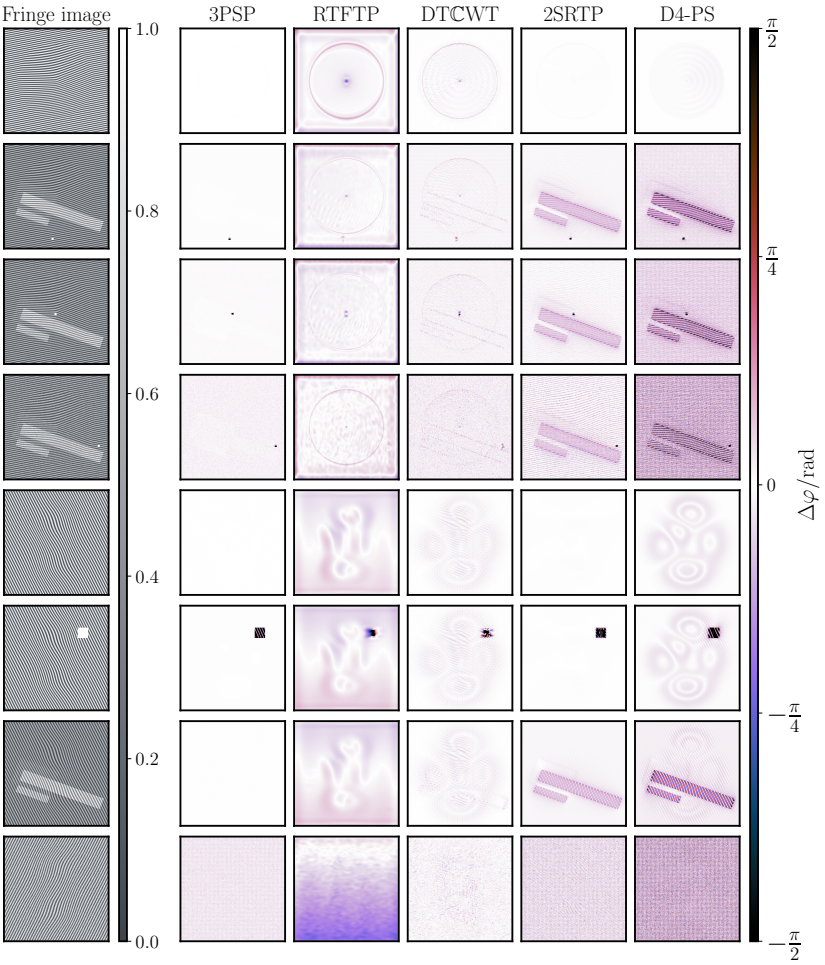


Figure 4: Comparison of reconstructed phase differences to the ground truth $\Delta\varphi$ of the cone object (1^{st} to 4^{th} row) and the peaks object (5^{th} to 8^{th} row) with different levels of noise and bias. The fringes on the cone are simulated at $\mathbf{f_0} = (-0.01, 0.05)\,\text{px}^{-1}$ and $\mathbf{f_0} = (0.04, 0.02)\,\text{px}^{-1}$ on the peaks function. The types of artifacts for each row are listed in Table 1.

3 Qualitative comparison

Table 2: Comparison between the presented methods in terms of required number of images versus computational time complexity for an image I of size $N \times N$.

Method	3PSP	RTFTP	DTCWT	2SRTP	D4-PS
Number of images	3	1	1	2	1
Time complexity $\mathcal{O}(\cdot)$	N^2	$N^2 \log(N)$	$N^2 \log(N)$	$N^2 \log(N)$	$N^2 \log(N)$

As mentioned above, even though the 3PSP method works best for reconstruction, the requirement of three images gives rise to the need of single- and few-shot alternatives in dynamic settings (see Tab. 2).

The automation of the commonly used FTP as a single-shot method with the RaT makes the framework suited for dynamic settings regardless of fringe frequency and orientation, reducing the need for manual intervention and multiple adjustments for the correct determination of filter parameters. However, artifacts in the images, particularly specular reflections and the reduction in spatial resolution due to filtering, pose performance bottlenecks in terms of quality. Even though the initial low- and high-pass filtering with L and H is effective, it has been observed in some cases that f_0, before applying B, was over- or underestimated due to corruption by high-frequency components in the SFD [8] introduced by the artifacts, therefore making the filtering of the carrier frequency component more challenging, which is also the reason for its complex implementation. This happens more prominently in combined cases of very low fringe contrast, low carrier frequencies and substantial artifacts, suggesting a minimally required signal quality for proper reconstruction. The computational time complexity of the repeated RaT dictates the overall time complexity and performance of the RTFTP with $\mathcal{O}(N^3)$, which can be reduced to $\mathcal{O}(N^2 \log(N))$ (see Tab. 2) with fast implementations of the RaT, making it approximately as computationally complex as the classical FTP [9,10].

In spite of the good performance of the DTCWT among the single-shot alternatives, the algorithm requires rough knowledge of fringe image properties prior to reconstruction, which may change in dynamic settings. This includes f_0 and $\frac{\partial \varphi}{\partial x}$ for identifying and isolating the relevant wavelet sub-bands $l_{\min}$ and $l_{\max}$, σ_n for Eq. (1) as well as

optimal spatial and range filter parameters to the median and bilateral filters. Most of these can be estimated or optimized beforehand using dedicated algorithms or a prior RTFTP, which, however, bloats the framework and increases computation time. Moreover, the design of a fully parameter-free, adaptive bilateral filter is still subject to current research. The time complexity of the DTCWT itself is $\mathcal{O}\left(N^2\right)$ [11]. However, the phase reconstruction using the RiT with r_v applied in the SFD with a FT forces the time complexity to $\mathcal{O}\left(N^2 \log(N)\right)$.

Overall, the RTP presents itself as the most versatile compared to the 3PSP and RTFTP, being applicable as a two (2SRTP) and single-shot method (DTCWT and D4-PS), while being easy to implement. This comes with the disadvantage that the RTP is increasingly sensitive to components with higher spatial frequency than the fringe frequency due to the Bedrosian theorem [12], which can be seen in the reconstructions. The 2SRTP and D4-PS demonstrate that there is a trade-off in the required number of images to correct for the high-frequency components and the achievable robustness to these components. However, the 2SRTP only works best if the bias is truly eliminated from the fringe as with the DTCWT, which was not achieved in Fig. 4 due to the clipping of the simulated fringes described above. The time complexity of the convolution operation in the RiT (see Eq. 2) can be reduced from $\mathcal{O}\left(N^4\right)$ to $\mathcal{O}\left(N^2 \log(N)\right)$ by applying it in the SFD.

Acknowledgements

The work described in this paper was conducted in the framework of the Graduate School 2543 "Intraoperative Multi-Sensory Tissue Differentiation in Oncology" (Project ID 409474577) funded by the German Research Foundation (DFG – Deutsche Forschungsgemeinschaft).

The authors thank Michael Saliba for administrative support and Valerie Klockner for typesetting code. The authors appreciate the support by Joshua Trapp and Yujia Kong for sharing important literature.

Author contributions Conceptualization: ÖA; Data curation: ÖA, WH; Formal analysis: ÖA, WH; Funding acquisition: AH; Investigation: ÖA, ME, WH; Methodology: ÖA, WH, ME; Project Administration: ÖA; Resources: ÖA; Software: ÖA, WH, ME; Supervision: ÖA, AH; Validation: ÖA, ME, WH; Visualization: ÖA, WH, ME; Writing -

original draft: ÖA, WH, ME; Writing - review & editing: ÖA, AH.

The results were produced using the very same version of the algorithm described in this paper. Certain aspects of the original algorithm were simplified after initial development and, in the process, enhanced in robustness using large language models (Gemini & ChatGPT).

References

1. M. Takeda *et al.*, "Fourier transform profilometry for the automatic measurement of 3-D object shapes," *Appl. Opt.*, vol. 22, no. 24, pp. 3977–3982, 1983.
2. S. R. Krishnan *et al.*, "On the Selection of Optimum Savitzky-Golay Filters," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 380–391, 2013.
3. W. W.-L. Ng *et al.*, "Effective bias removal for fringe projection profilometry using the dual-tree complex wavelet transform," *Appl. Opt.*, vol. 51, no. 24, pp. 5909–5916, 2012.
4. W. W.-L. Ng *et al.*, "Image enhancement for fringe projection profilometry," in *2013 IEEE International Symposium on Circuits and Systems*. Beijing, China: IEEE, 2013, pp. 729–732.
5. M. Felsberg *et al.*, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, 2001.
6. Y. Tounsi *et al.*, "Single-shot phase-shifting interferometry: methods, challenges, and future perspectives," *Adv. Opt. Technol.*, vol. 15, 2026.
7. J. M. Bioucas-Dias *et al.*, "Phase Unwrapping via Graph Cuts," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 698–709, 2007.
8. G. Wetzstein *et al.*, "Sensor saturation in Fourier multiplexed imaging," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, 2010, pp. 545–552.
9. P. Duhamel *et al.*, "Fast fourier transforms: A tutorial review and a state of the art," *Signal Process.*, vol. 19, no. 4, pp. 259–299, 1990.
10. M. L. Brady, "A Fast Discrete Approximation Algorithm for the Radon Transform," *SIAM J. Comput.*, vol. 27, no. 1, pp. 107–119, 1998.
11. D. Wei *et al.*, "Wavelet Denoising for Image Enhancement," in *Handbook of image and video processing*, A. C. Bovik, Ed. Academic Press, 2000, ch. 3.4, p. 118.
12. E. Bedrosian, "A product theorem for Hilbert transforms," *Proc. IEEE*, vol. 51, no. 5, pp. 868–869, 1963.

Kombination der Stärken von Simulation und generativer KI zur synthetischen Bildgenerierung

Combining the strengths of simulation and generative AI for synthetic image generation

Anne Sielemann[1], Masoud Roschani[1],
Jens Ziehn[1] und Jürgen Beyerer[12]

[1] Fraunhofer IOSB, Fraunhofer Straße 1, 76131 Karlsruhe
[2] Karlsruher Institut für Technologie (KIT),
Lehrstuhl für Interaktive Echtzeitsysteme (IES), 76131 Karlsruhe

Zusammenfassung Die Bildsynthese mittels analytischer oder datengetriebener Modelle weist jeweils individuelle Vor- und Nachteile auf. Mit hybriden Ansätzen wird versucht, die Stärken beider Ansätze zu kombinieren. Im Rahmen dieser Arbeit stellen wir unsere Rendering-Pipeline der Simulationsplattform OCTAS® und insbesondere deren Integration von generativer KI vor. Die Pipeline implementiert das Prinzip des physikalisch-basierten Renderns und beinhaltet Schnittstellen zu rasterisierungsbasierten und path-tracing-basierten Rendering Engines. Ferner enthält die Pipeline ein umfangreiches Kameramodell, um die Bilder aus dem analogen in den digitalen Bildraum zu überführen. Anhand verschiedener konkreter Anwendungsfälle erläutern wir, wie die hybriden Ansätze konkret realisiert werden können und messen den entsprechenden Effekt auf die Performanz Metriken der jeweiligen Aufgabe. Des weiteren diskutiert die Arbeit, wie hybride Synthetisierungsansätze auf weitere Anwendungsfälle übertragen werden können und was dabei zu berücksichtigen ist.

Schlüsselwörter Simulation, Bildsynthese, generative KI

Abstract Image synthesis using analytical or data-driven models each has its own specific advantages as well as disadvantages.

Hybrid approaches aim to combine the strengths of both. In this work, we present our rendering pipeline of the OCTAS® simulation platform and, in particular, its integration of generative AI. The pipeline implements the principles of physically based rendering and includes interfaces to both rasterization-based and path-tracing-based rendering engines. Furthermore, the pipeline contains an extensive camera model to transfer images from the analog to the digital image space. Using several specific application scenarios, we explain how hybrid approaches can be realized in practice and measure their effect on the performance metrics of the respective task. In addition, this work discusses how hybrid synthesis approaches can be transferred to further application scenarios and what must be taken into account in doing so.

Keywords Simulation, image synthesis, generative AI

1 Einleitung

Synthetische Bildgenerierung gewinnt zunehmend an Bedeutung, insbesondere zur Skalierung und Diversifizierung von Trainingsdaten für Anwendungen des maschinellen Lernens im Bereich Computer Vision – mit dem perspektivischen Ziel des vollumfänglichen Trainierens und Testens komplexer und ggf. sicherheitskritischer Systeme wie beispielsweise automatisierte Fahrzeuge. Zwei grundlegende Synthetisierungsansätze stehen hierbei im Vordergrund: generative KI-Modelle und physikalisch basierte Simulation bzw. 3D-Rendering. Beide verfolgen das Ziel, realitätsnahe Bilddaten bereitzustellen, unterscheiden sich jedoch hinsichtlich Eigenschaften wie Realismus, Variation, Datenbedarf, Kontrollierbarkeit und Annotationstreue.

Generative KI-Modelle lernen automatisiert und datengetrieben komplexe Korrelationen und hochdimensionale Verteilungen von Bildmerkmalen. Zu ihren zentralen Vorteilen zählen eine gute Annäherung an natürliche Bildstatistiken und ein daraus resultierender hoher Realismus bei geringen manuellen Aufwänden, allerdings um den Preis einer begrenzten Kontrollierbarkeit und ungenaueren Ground-Truth-Annotationen. Zudem sind generative Modelle von der Verfügbarkeit großer, qualitativ hochwertiger Trainingsdatensätze abhängig, die die

Zieldomäne hinreichend gut abdecken müssen und reproduzieren ggf. bestehende Verzerrungen (Bias) aus diesen Daten.

Demgegenüber stehen simulationsbasierte Verfahren, die die Bildentstehung vollumfänglich modellieren: 3D-Geometrie, Materialien, Lichtquellen, Sensorcharakteristika und physikalische Interaktionen sind explizit modelliert, parametrierbar und somit kontrollierbar. Dies ermöglicht Reproduzierbarkeit, die systematische Erzeugung seltener oder gefährlicher Szenarien, die Erweiterbarkeit von Datensätzen ohne Domänenunterschied sowie die automatische Generierung exakter Annotationen (z.B. semantische Segmentierung, Tiefeninformation). Es geht jedoch mit Aufwand für die Modellierung sowie einem potenziellen systematischen Mangel an Realismus (der sogenannte *„sim-to-real-Gap"*) einher, der die Übertragbarkeit auf reale Daten einschränkt.

Hybride Ansätze bieten das Potenzial, die Stärken beider Methoden zu kombinieren sowie deren jeweilige Schwächen abzumildern. Wir stellen in dieser Arbeit vor, wie eine Kombination beider Ansätze dazu beitragen kann, eine physikalisch korrekte Darstellung von analytisch modellierbaren Parametern zu erhalten, die auf physikalische Zusammenhänge in der Bilderzeugung zurückschließen lassen, und somit auch kritische Aspekte eines physischen Realsystems abschätzbar machen – und gleichzeitig komplexe Umgebungseigenschaften über moderne KI-Verfahren kontrollierbar abgebildet werden können. So entsteht eine Bilderzeugungskette für synthetische Daten, die sowohl variantenreich und flexibel in Bezug auf komplexe Open-World-Anwendungen ist, als auch physische Konzepte konsequent und steuerbar modelliert.

2 Stand der Technik

2.1 Generative KI zur Bildsynthese

Frühe Ansätze von generativer KI zur Bildsynthese basierten überwiegend auf *Variational Autoencoders* (VAEs) [1] und *Generative Adversarial Networks* (GANs) [2]. VAEs ermöglichen die Abbildung komplexer Bilddaten in kontinuierliche latente Räume, weisen jedoch häufig Einschränkungen in der Detailtreue auf. GAN-basierte Verfahren erzielen demgegenüber erhebliche Fortschritte hinsichtlich Fotorealismus und Auflösung, sind jedoch mit Herausforderungen wie Trai-

ningsinstabilität, Moduskollaps und begrenzter Kontrollierbarkeit verbunden. Gegenwärtig dominieren *Diffusionsmodelle* [3,4] den Stand der Technik. Diese modellieren die Bildgenerierung als iterativen Entrauschungsprozess und vereinen hohe Synthesequalität mit vergleichsweise stabilen Trainingsdynamiken. Insbesondere Latente Diffusionsmodelle [5] haben durch die Verlagerung des Diffusionsprozesses in den latenten Raum eine signifikante Effizienzsteigerung ermöglicht. Die Generierung der Bilder kann beispielsweise durch die Eingabe von Referenzbildern, Text, semantischen Masken oder Skizzen gesteuert werden. Einen tiefergehenden Einblick zu den genannten und weiteren Ansätzen gibt [6].

2.2 3D-Rendering und Simulationsansätze zur Bildsynthese

Zu den Ansätzen der simulationsbasierten Bildsynthese zählen einerseits Ansätze, die direkt Computerspiel-Engines wie *Unity*[3], *Unreal*[4] oder *Godot*[5], eine 3D-Grafiksoftware wie *Blender*[6] oder die API eines Renders verwenden. Diese Ansätze verbindet, dass sie sehr flexibel genutzt und für beliebige Anwendungsfälle angepasst werden können. Erforderlich ist jedoch ein hoher Modellierungsaufwand, da Logik implementiert werden muss, um gewünschte Szenarien aufzubauen und zu variieren. Bibliotheken wie beispielsweise *BlenderProc2* [7] können im Fall von Blender dabei unterstützen, die Komplexität zu abstrahieren. Deutlich geringerer Modellierungsaufwand besteht andererseits bei der direkten Nutzung von Computerspielen zur Bildsynthese, wie beispielsweise in [8, 9]. Diese lassen sich allerdings gar nicht oder kaum an individuelle Anwendungsfälle anpassen, und die Extraktion der Ground Truth-Daten kann eine Herausforderung darstellen. Einen Kompromiss bieten Simulatoren wie CARLA [10] oder OCTAS® [11]: im Allgemeinen vereinfachen sie den Modellierungsaufwand, sind aber gleichzeitig explizit dafür ausgelegt, individuelle Anwendungsszenarien zu simulieren.

[3] `https://unity.com`
[4] `https://www.unrealengine.com/`
[5] `https://godotengine.org`
[6] `https://www.blender.org/`

2.3 Hybride Ansätze

Das Ziel hybrider Ansätze, die Simulation und generative KI miteinander verbinden, besteht in der Regel darin, die Parametrierbarkeit und damit die Kontrollierbarkeit simulationsbasierter Verfahren mit dem visuellen Realismus und der Varianz von generativer KI zu vereinen. Hierzu werden typischerweise entweder einzelne Komponenten des simulationsbasierten Syntheseprozesses durch generative KI ersetzt oder die Prozesskette um eine KI-gestützte Nachbearbeitung erweitert. Im erstgenannten Fall kann beispielsweise die Texturgenerierung für 3D-Geometrien durch generative KI substituiert werden [12]. Darüber hinaus wird intensiv daran geforscht, KI in den Renderingprozess zu integrieren. Dabei werden aus Bilddaten Szenenrepräsentationen gelernt, aus denen anschließend neue Ansichten synthetisiert werden können. Beispiele hierfür sind *Neural Radiance Fields*, kurz *NeRF* [13], oder *3D Gaussian Splatting* [14]. Zur Nachbearbeitung von Simulationsbildern durch generative KI ist es insbesondere populär, neuronalen *Style Transfer* [15] anzuwenden, um die generierten Daten der Zieldomäne stärker anzugleichen. Zudem kann die Datenbasis mittels neuronaler *Inpaintings* [16] erweitert oder einzelne Bildeigenschaften gezielt mittels sogenanntem *Attribute Editing* verändert werden [17].

3 Methodik

Auf abstraktem Level besteht die OCTAS® Rendering Pipeline zur Generierung von Sensordaten aus vier Bausteinen: (1) dem Aufbau des 3D-Szenarios im Objektraum, (2) der Überführung der Szene vom Objektraum in den (analogen) Bildraum, (3) der Simulation von Kameraeffekten im analogen Bildraum und (4) der Überführung vom analogen in den digitalen Bildraum durch Diskretisierung und Simulation von Digitalisierungseffekten. Ein schematischer Überblick mit Darstellung der Unterschritte jedes dieser Pipeline-Bausteine ist in Abbildung 1 gegeben. Das Design der Pipeline folgt insbesondere zwei leitenden Designprinzipien: Modularität und physikalisch basiertes Rendern, engl. *Physically Based Rendering* (PBR) [18]. Das Ziel von PBR ist dabei, den gesamten Renderingprozess, insbesondere die Wechselwirkungen von Licht und Materialien, mittels physikalisch plausibler Modelle so abzubilden, dass realistische, konsistente und szenenübergreifend vor-

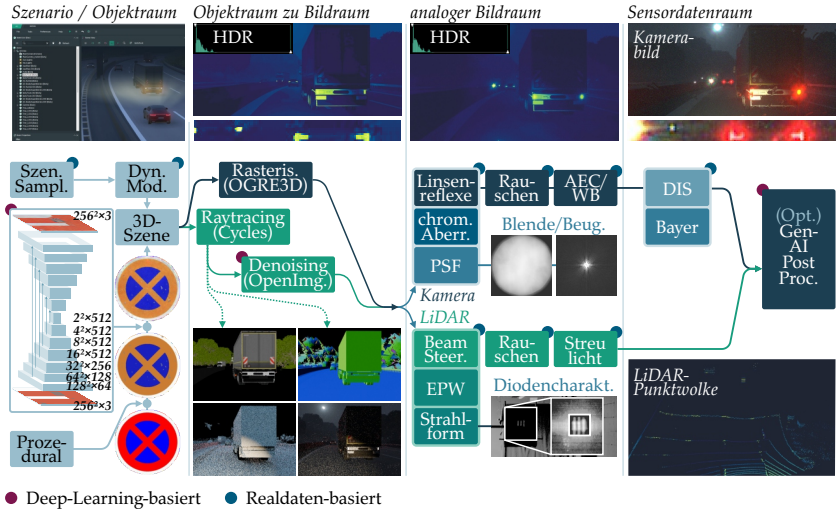


Abbildung 1: Schematische Darstelltung der Rendering Pipeline von OCTAS®.

hersagbare Bildresultate entstehen. Auf diese Weise unterstützt PBR das Designprinzip der Modularität, da sich Komponenten wie etwa die eingesetzte Rendering-Engine dank der standardisierten Darstellung von Materialien, Lichtern etc. einfach und dennoch vergleichbar austauschen lassen.

3.1 Prinzipien der Integration von generativer KI

Die beiden leitenden Designprinzipien finden auch bei der Integration von generativer KI in die OCTAS® Rendering Pipeline Anwendung. Während die Modularität es prinzipiell erlaubt, jede Komponente auszutauschen und durch ein KI-Pendant zu ersetzen, erfordert PBR eine physikalisch basierte Modellierung und Bildberechnung. Allerdings kann die physikalische Korrektheit der Beleuchtung durch generative KI nur bedingt gewährleistet werden, da Beleuchtung aus dem komplexen Zusammenspiel von Geometrie, Material, Lichtquelle und Sichtwinkel hervorgeht. Aktuelle bildgenerierende Modelle erlernen in der Regel primär statistische Korrelationen aus zweidimen-

sionalen Bilddaten, verfügen jedoch nicht über ein vollständiges physikalisches Verständnis der zugrunde liegenden 3D-Szene. Aus diesem Grund wird in der OCTAS® Rendering-Pipeline derzeit vermieden, Aufgaben wie die Beleuchtungsberechnung oder die umfassende Darstellung von Materialeigenschaften durch generative KI zu ersetzen. Ebenso werden keine ausschließlich KI-generierten Objekte mittels neuronaler *Inpaintings* nachträglich in die Bilder eingefügt.

Bei Texturen stellt sich die Situation hingegen anders dar. Diese lassen sich nur mit erheblichem Aufwand analytisch modellieren, da sie durch hohe Variabilität, Unregelmäßigkeiten und zahlreiche lokale Abweichungen gekennzeichnet sind. Sie folgen daher in der Regel keinen einfachen, geschlossenen mathematischen Gesetzmäßigkeiten, sondern entstehen aus komplexen stochastischen Prozessen. Eine manuelle Modellierung und Parametrisierung ist entsprechend aufwendig, insbesondere wenn realitätsnahe Ergebnisse erzielt werden sollen. Generative KI kann Textureigenschaften hingegen vergleichsweise zuverlässig lernen, da diese überwiegend durch lokale und statistisch wiederkehrende Bildmuster charakterisiert sind. Voraussetzung hierfür ist jedoch, dass die Trainingsdaten eine ausreichende Variationsbreite aufweisen und möglichst frei von eingebackenen Beleuchtungsartefakten sind. Nur so kann eine konsistente Verwendung der generierten Texturen unter unterschiedlichen Beleuchtungsszenarien gewährleistet werden. Darüber hinaus sollten die resultierenden Texturen möglichst PBR-tauglich sein. Das bedeutet insbesondere, dass ergänzende Texturen, beispielsweise für Normalen, Rauheit oder Metallizität, entweder gemeinsam oder nachträglich generiert werden können.

Eine Zwischenposition nimmt der *Style Transfer* ein, der sowohl zur *Domain Adaptation* als auch zur Steigerung des Realismus simulierter Bilder eingesetzt werden kann. Da sich die generative KI hierbei an den von der Simulation vorberechneten Szenenstrukturen und Beleuchtungsverhältnissen orientieren kann, bleiben wesentliche globale Strukturen, insbesondere Schattenwürfe und Lichtrichtungen, in der Regel annähernd erhalten. Dadurch sind die Ergebnisse tendenziell realistischer, als dies bei einer vollständigen Generierung vergleichbarer Bilder durch KI der Fall wäre. Jedoch können lokale Strukturen wie Objektformen und -details, Schattengrenzen, Reflexionen und Glanzlichter von der KI modifiziert werden. Zudem überträgt der Style Transfer in der Regel auch die Farbverteilung und den Kontrast, wo-



Abbildung 2: Beispiele für mit der OCTAS® Rendering Pipeline generierte synthetische Bilder zur Verkehrszeichenklassifikation mit einfachen Templates als Texturen (obere Reihe) und durch ein GAN variierte Texturen (untere Reihe).

durch die physikalische Korrektheit in der Regel nicht gewährleistet ist. Für Anwendungen, die hohe Anforderungen an die physikalische Korrektheit der Daten stellen, ist die Erzeugung von Trainingsdaten mittels Style Transfer daher nicht zu empfehlen bzw. nur mit sorgfältiger Prüfung durchzuführen. Zur Domain Adaptation kann dieses Verfahren hingegen ein geeignetes und wirkungsvolles Werkzeug darstellen.

4 Anwendungsbeispiele

4.1 Daten zur Verkehrszeichenklassifikation

Verkehrszeichen sind einer Vielzahl externer Einflüsse ausgesetzt, wie beispielsweise Sonnenlicht, Witterung, Temperaturwechsel oder Vandalismus. Diese verursachen unter anderem das Ausbleichen der Farbe, Risse in der Oberfläche, Ablagerungen von Dreck oder Verformungen, die die Oberfläche jeder einzelnen Verkehrszeicheninstanz innerhalb einer Klasse unterschiedlich aussehen lassen. Denen im vorherigen Abschnitt vorgestellten Prinzipien folgend, sollen diese Effekte berücksichtigt werden, indem die individuellen Verkehrszeichentexturen mittels generativer KI erzeugt werden. Dabei nutzen wir ein Pix2Pix GAN [19], um realistische Verkehrszeichenmerkmale auf prozedural erzeugte Templates zu übertragen, ohne deren Struktur zu verändern, wodurch der Grad der Degradierung parametrierbar bleibt (vgl. Abbildung 1). Mehr Informationen zum Generierungsprozess in [20,21]. Um den Effekt auf die Performance zu messen, vergleichen wir die Klassifikationsergebnisse von zwei *ConvNext V2* Netzwerken [22],



Abbildung 3: Beispiele des Style Transfers von SSB [25] in die Domäne von CCSV [26] mittels drei verschiedener GAN Architekturen. Quelle: [27].

die auf zwei synthetischen Datensätzen mit jeweils 198 Bildern für jede der 43 enthaltenen Klassen trainiert wurden. Einer der Datensätze (A) nutzt mittels generativer KI erzeugte Texturen, der andere Datensatz (B) verwendet unmodifizierte Originalbilder der Verkehrszeichen zur Texturierung. Bis auf diesen Unterschied wurden beide Datensätze identisch parametriert. Abbildung 2 enthält Beispielbilder beider Datensätze. Als Validierungsdatensatz wird *GTSRB* [23] verwendet, unter Nutzung des Evaluationskonfiguration aus [24]. Gemittelt über drei Trainingsläufe mit unterschiedlichem Seed erreicht Netz (B) eine Accuracy von $0,878 \pm 0,001$. Netz (A) übertrifft diese um 0,026 und erreicht eine Accuracy von $0,904 \pm 0,007$.

4.2 Daten zur Fahrzeugmarken- und Modellerkennung (VMMR)

Die Arbeit von Krohmer et al. [27] zeigt am Beispiel der feingranularen Klassifikation von Fahrzeugmarken und -modellen (VMMR), wo die Vorteile, aber auch die Herausforderungen bei einem Style Transfer von synthetischen Daten in eine andere Zieldomäne liegen. Dabei transferieren die Autoren den mit der OCTAS® Rendering Pipeline generierten *Synset Boulevard* Datensatz (SSB) [25] in die Domäne des realen *CompCars Surveillance* Datensatzes (CCSV) [26]. Beide Domänen unterscheiden sich nicht nur durch den Realismusgrad bzw. die Erzeugungsquelle, sondern auch durch einen geografischen Bias. Abbildung 3 zeigt

Beispielresultate. Durch die Veränderung von diskriminativen Merkmalen wie beispielsweise des Markenlogos erzielen Ansätze wie CycleGAN eine Verbesserung des F1-Scores um 0,122 bei der Evaluation auf CCSV [27]. Ansätzen wie dem vorgeschlagenen FCCGAN [27], die darauf abzielen, diskriminative Features zu erhalten, gelingt sogar eine Verbesserung des F1-Scores um 0,286.

5 Zusammenfassung und Ausblick

Die Kombination von physikalisch basierter Simulation und generativer KI verbindet die Stärken beider Einzelansätze und bietet somit eine vielversprechende Grundlage für realistische, variantenreiche und zugleich kontrollierbare synthetische Datenerzeugung. Die OCTAS® Rendering-Pipeline demonstriert, wie das Framework des physikalisch basierten Renderns mit generativer KI kombiniert werden kann, sodass physikalisch modellierbare Eigenschaften wie Geometrie, Materialeigenschaften, Beleuchtung und Kameraparameter kontrolliert bleiben, während komplexe Texturen und Domänenmerkmale datengestützt ergänzt werden. Die Anwendungsbeispiele zeigen, dass dadurch sowohl die Performance als auch die Übertragbarkeit synthetischer Daten verbessert werden können. Zukünftige Forschung sollte die physikalische Plausibilität der erzeugten Inhalte weiter erhöhen und zugleich effiziente Methoden zur automatisierten Qualitätsbewertung entwickeln.

Literatur

1. D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
2. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” *Advances in neural information processing systems*, vol. 27, 2014.
3. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.
4. J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

5. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
6. J. Li, C. Zhang, W. Zhu, and Y. Ren, "A Comprehensive Survey of Image Generation Models Based on Deep Learning," *Annals of Data Science*, vol. 12, no. 1, pp. 141–170, 2025.
7. M. Denninger, D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. W. Knauer, K. H. Strobl, M. Humt, and R. Triebel, "BlenderProc2: A Procedural Pipeline for Photorealistic Rendering," *Journal of Open Source Software*, vol. 8, no. 82, p. 4901, 2023.
8. M. Johnson-Roberson, C. Barto, R. Mehta *et al.*, "Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks?" *arXiv preprint arXiv:1610.01983*, 2016, sridhar, Sharath Nittur and Rosaen, Karl and Vasudevan, Ram.
9. M. Scucchia, P. Arranz, M. Ferrara, and D. Maltoni, "From Gaming to Research: GTA V for Synthetic Data Generation for Robotics and Navigations," in *2025 7th International Conference on Robotics and Computer Vision (ICRCV)*. IEEE, 2025, pp. 187–196.
10. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.
11. J. Ziehn, M. Roschani, A. Sielemann *et al.*, "OCTAS," https://octas.org/, 2025.
12. A. Hu, N. Desai, H. Abu Alhaija, S. W. Kim, and M. Shugrina, "Diffusion Texture Painting," in *ACM SIGGRAPH 2024 Conference Papers*, 2024.
13. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
14. B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis *et al.*, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
15. L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv preprint arXiv:1508.06576*, 2015.
16. W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka, "Deep Learning-Based Image and Video Inpainting: A Survey," *International Journal of Computer Vision*, vol. 132, no. 7, pp. 2367–2400, 2024.

17. X. Li, Y. Chen, Y. Zhu, S. Wang, R. Zhang, and H. Xue, "ImageNet-E: Benchmarking Neural Network Robustness via Attribute Editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 371–20 381.
18. M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation*. MIT Press, 2023.
19. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
20. A. Sielemann, L. Loercher, M.-L. Schumacher, S. Wolf, M. Roschani, J. Ziehn, and J. Beyerer, "Synset Signset Germany: a Synthetic Dataset for German Traffic Sign Recognition," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2024, pp. 3383–3390.
21. A. Sielemann, "The Synset Signset Rendering Pipeline – A Principle for Generating Synthetic Data for ML," in *Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*, 2026.
22. S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.
23. J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.
24. A. Sielemann, N. Bitzer, M. Roschani, J. Ziehn, and J. Beyerer, "Sensitivity Analysis of Deep-Learning Perception Models to Sensing Parameters," in *2026 IEEE 29th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2026.
25. A. Sielemann, S. Wolf, M. Roschani, J. Ziehn, and J. Beyerer, "Synset Boulevard: A Synthetic Image Dataset for VMMR," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9146–9153.
26. L. Yang, P. Luo *et al.*, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, change Loy, Chen and Tang, Xiaoou.
27. E. Krohmer, S. Wolf, and J. Beyerer, "Supervised domain adaptation with disjoint label spaces for fine-grained classification," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 50–66.

Precise monocular ball localization in 4K through agentically generated synthetic training data

Tobias Pruß, Jonas Breuling,
Tianxiang Dai, and Remco I. Leine

University of Stuttgart, Institute for Nonlinear Mechanics,
Pfaffenwaldring 9, 70569 Stuttgart, Germany

Abstract An objective measurement of spin in recreational table tennis using a single smartphone camera and unmarked standard balls may provide valuable training feedback. To the authors' knowledge, the spin estimated by markerless single-camera approaches in the literature has so far been validated only qualitatively. A quantitative analysis presupposes a precisely localized ball centre. Common trackers, however, operate at reduced resolution and are assessed against hand-annotated data. The goal of this work is subpixel-accurate ball localization on native 4K footage. In the method proposed here, existing trackers trained on real data provide a robust coarse localization. To supply the missing precision, a downstream neural refiner is introduced which evaluates a patch around the coarse position at native resolution. The refiner is trained exclusively on synthetic data with analytically exact ground truth, generated largely automatically from the rules of the International Table Tennis Federation (ITTF) by coding agents based on large language models (LLMs). A mechanical reference rig with caliper-measured geometry makes selected parts of the sim-to-real error measurable in a controlled setting. Using this reference as feedback, the agents autonomously revise the synthetic training data and reduce the measured errors to a fraction of a pixel. On rally footage recorded in a sports hall with two smartphones, the refined centres cut the scatter of the triangulated ball positions from 5 mm to 1 mm.

Keywords Ball localization, synthetic training data, physical reference, agentic workflow, table tennis

1 Introduction

The aim of this paper is to locate the centre of an unmarked table tennis ball in native 4K smartphone footage to a fraction of a pixel, as the foundation for measuring spin with a single phone. Its contribution is a method in which LLM-based coding agents generate the synthetic training data for this task and revise it against a physical reference measurement, so that the achievable precision is no longer limited by the accuracy of hand-set labels.

In table tennis, spin governs the flight and the bounce of the ball, and it is among the quantities hardest to judge in play. Players infer the opponent's spin from the stroke and judge their own by feel at the racket contact, which yields no quantitative value. An objective measurement with a single smartphone camera and unmarked standard balls, and thus with equipment that is readily available, would provide exactly this missing feedback.

Reading the spin off the ball itself has been demonstrated with a dot pattern drawn on the ball [1] or the manufacturer's logo, together with high-speed or multi-camera set-ups [2,3]. The commercial system *Spinsight* measures spin with a smartphone camera but also relies on a marked ball [4]. Markerless methods with a single camera infer spin from the trajectory [5–7], but on real footage the spin they estimate has so far been validated only qualitatively. Spin acts through the curvature of the trajectory and through the bounce, both small effects that are read from the ball centre. The precision of that centre therefore sets the precision any spin measurement from the trajectory can reach.

The ball centre is supplied by a tracker. Trackers of the TrackNet family [8], WASB [9] and BlurBall [10] locate the ball reliably, but the lowest mean distance error reported on a table tennis benchmark is 3.4 px, for a TrackNet variant evaluated against hand-set labels [11], which is not precise enough for an accurate quantitative spin measurement. The precision of those trackers is bounded in two ways. The trackers process the frame at a reduced resolution [8–10], and they learn from hand-annotated datasets [8–11]. A network trained on hand-set labels tends to be only as accurate as those labels.

In this work a method is proposed to locate the ball centre at full resolution more precisely than a hand-set label allows, while keeping the robustness of the trackers. To this end, a second network is placed

behind the tracker, a refiner that works on a patch of the frame at full resolution and is trained on synthetic data with analytically exact ground truth. The development of the refiner is realized in three steps, large parts of which LLM-based coding agents carried out on their own.

1. Train the refiner on synthetic data alone, with analytically exact ground truth.
2. Test the refiner against real reference measurements and use them to improve it.
3. Test the resulting pipeline on real recordings.

Section 2 describes the refiner and its synthetic training data. Section 3 presents the reference measurement on the rig and the revisions of the training data that followed from it. Section 4 tests the pipeline on real rallies. Section 5 describes the agentic workflow, and Section 6 concludes with the limits and the outlook.

2 Synthetic Training Data

The tracker, in this work BlurBall [10], serves as the coarse detector and supplies the position around which the refiner evaluates a 64×64 patch of the native 4K frame. The training data of the refiner is purely synthetic. LLM-based coding agents built the generator largely on their own from the dimensions of ball and table laid down by the International Table Tennis Federation (ITTF) [12], rendered the training patches with analytically exact ground truth and trained the first refiner on them. On a synthetic test set from the same generator this refiner reached a median error below half a pixel.

The accuracy of the refiner trained on synthetic data says nothing about its performance on real recordings. Synthetic test data shares every assumption of the synthetic training data, so it cannot reveal where the renderer differs from the images taken by a real camera. A hand-set label on real footage cannot reveal that difference either, since the label is coarser than the quantity to be measured. Judging the refiner therefore requires a reference measurement on real footage that is more accurate than a hand-set label.

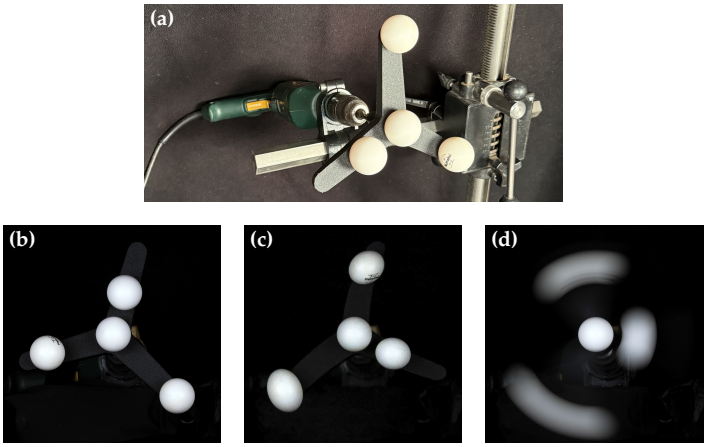


Figure 1: The reference rig. (a) The single-piece printed part with the four glued-in regulation balls in the chuck of a hand drill on a drill stand. (b) At rest at 1/125 s, (c) rotating at 1/32000 s and (d) rotating at 1/125 s, photographed with a mirrorless camera. These frames illustrate motion blur and sensor readout.

3 Reference Measurement

A reference measurement for the refiner needs real footage from the target device, since the method is meant to run in a smartphone app. Furthermore, the ball centres in that footage need to be determined more accurately than manual labelling can. The conditions have to be variable one at a time, from a ball at rest under changing light, exposure and focus to one moving at the speed of play. Each disturbance can then be measured on its own and the refiner tested against it. A 3D-printed rig provides these conditions.

The rig is a single-piece part that carries four unmarked regulation balls, three at orbital radii of 49.77, 78.01 and 101.45 mm and one on the axis of rotation, and sits in the chuck of a hand drill in front of a black backdrop (Fig. 1a). Its geometry was measured with a digital caliper, at the bore holes before the balls were glued in and over the balls afterwards. At rest the rig is a ball holder of known geometry. Turning, it moves the balls at nearly constant speed around a fixed position, so that motion blur and sensor readout occur under controlled

conditions. An iPhone 16 Pro on a tripod about 3 m from the rig records 4K video at 60 frames per second with the camera automatics switched off. At this distance the ball measures about 30 px in diameter, its typical size in a rally. The recording format is ProRes 422 HQ with Apple Log, the least compressed format the phone offers.

At 30 px the effects of the ball speed on motion blur and on the distortion of the apparent shape by the sensor readout are hard to see. To illustrate them, the rig was also photographed with a mirrorless camera, stabilization off and focus set by hand, which frames the balls at about 360 px (Fig. 1b–d). At 1/32000 s the motion is practically frozen, yet the moving balls are drawn into ellipses by the line-by-line readout of the sensor while the ball on the axis stays round. At 1/125 s every ball draws an arc whose length grows with its orbital radius.

The authors defined the outline of the measurement campaign, a series at rest under varied gain, exposure time, defocus and light direction and a series with the rig turning. Neither the phone nor the drill has been designed as a measuring instrument. The drill does not run perfectly true and its speed is not perfectly constant, and stepping the exposure time, gain, focus or white balance of the phone during a recording, with every setting logged, requires a custom capture app.

The agents built the capture app, planned the details of the campaign and ran it. The authors' part was limited to plugging in the phone, switching on the light and starting the drill. Afterwards the agents fetched the recordings from the phone, filed them and documented every setting. They measured and modelled the disturbances the authors knew of. The readout time of the sensor was found to be 2.28 ± 0.22 ms for the full frame, in line with the 2.4 ms that slashCAM measured on the iPhone 16 Pro Max [13]. The agents determined the run-out of the drill to be below 0.1 mm and its speed to be 1517 revolutions per minute with a scatter of 0.3 %. They also characterized the direction of the light and the saturation of the lit side of the ball. The agents even found effects the authors had not anticipated. According to their measurements the phone drifts on its tripod by 2.6 px over a session. Its optical image stabilization cannot be switched off for video through the public API, so the image keeps moving during a recording. They also observed that the focal length reported by the phone drifts over the minutes of a recording even with the focus locked.

With the measured disturbances accounted for, and with a printed



(a) Synthetic, from the ITTF dimensions

(b) Real rig recordings

(c) Synthetic, after calibration

Figure 2: Training patches of the first generator (a) and of the final revision (c) beside real patches from the rig recordings (b), shown once in colour and once as the luminance channel alone.

ChArUco calibration board [14] fixed beside the rig as the stationary reference in the image, the agents then built the reference centres from the recordings. At rest, the ball centres were measured on sharp and well-exposed recordings at the start and the end of the measurement campaign with three classical methods implemented independently by the agents. The three methods are an edge fit on the silhouette, a robust circle fit on radial gradients and a forward model of the blurred disc, each estimating the ball centre from a different feature of the image. A ball centre was accepted only where all three methods agreed to within a tenth of a pixel, and the accepted centres serve as the reference for the whole series at rest. With the rig turning, the centres come from a rigid-body model fitted per frame and anchored on the axis ball, so that the run-out of the drill is accounted for.

Against these reference centres the agents then revised the training data, making amendments based on the footage itself. Each of the disturbances measured above was given an uncertainty and included in the renderer. The authors checked these measurements for plausibility, for instance the readout time against the published value, but did not verify each step of the agents’ work individually.

What justifies confidence in the result is the role of the synthetic data. Whatever the agents conclude about the appearance of the ball or the behaviour of the hardware is written into the renderer as a parameter, never into a label. In the renderer the parameter is analytically



Figure 3: The sports-hall set-up. Two smartphones on light stands beside the table, the laptop with the capture app between them, and printed ChArUco boards on the table against which each camera is calibrated independently.

exact and can be varied independently of every other parameter. Each conclusion thus predicts where the refiner's error against the reference centres should fall, and a conclusion whose prediction fails is revised in the next iteration. A mistaken conclusion therefore does not hide in the training data but appears as a failed prediction against the reference.

The first refiner, built on the ITTF dimensions, reached a median error of 1.68 px at rest and 2.26 px turning. After the revisions the median error was 0.13 px at rest and 0.27 px turning. Figure 2 shows the training data before and after these revisions beside patches from the real reference recordings.

4 Real Rallies in a Sports Hall

To test the pipeline in real play, the authors recorded rallies in a sports hall with two smartphones, an iPhone 16 Pro and an iPhone 12 mini, both at 4K and 60 frames per second with locked settings (Fig. 3). Printed ChArUco boards on the table give the position of the two cameras relative to each other and to the table. The refiner runs on both recordings behind the coarse detector. The top row of Fig. 4 shows patches from a serve with the raw mark of the coarse detector and the refined centre. The coarse detector works on the frame scaled down to 512×288 pixels [10], so one of its pixels covers 7.5 pixels of the 4K

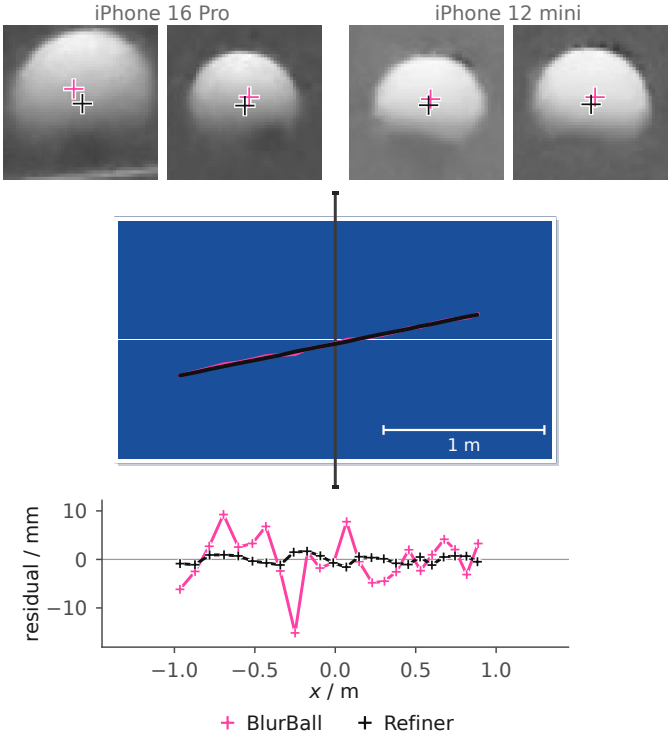


Figure 4: Coarse detector against the refiner on a low-spin serve. Top, patches from two frames of each camera with the raw mark of the coarse detector and the refined centre. Middle, the flight between the two bounces on the table, every frame pair triangulated on its own with the camera geometry held fixed. Bottom, the lateral residual of each track against its own straight-line fit.

frame. Its mark falls several pixels off the centre of the ball. The refined mark lies close to the centre.

The two phones are not triggered in hardware, and no independent reference for the ball position is available in the hall, so an exact ground truth for a moving ball cannot be obtained with this set-up. What can be measured is how much the reconstructed positions scatter, since an error common to all frames, such as a constant timing offset between the two cameras, shifts all positions by the same amount and does not

show in the scatter. For this measurement the authors selected a serve with little spin, whose flight between the two bounces comes close to a straight line in the top view, because gravity acts vertically and the spin-induced Magnus force is small. The two recordings were synchronized by the sound of the ball contacts on both audio tracks. Every frame pair was then triangulated on its own, and the scatter of each detector is measured against a straight line fitted through its triangulated positions. The positions of the coarse detector deviate by up to 15 mm from their line, with a root mean square of 5 mm, almost entirely along the depth direction of the camera pair, the direction a stereo pair constrains least. The refined positions stay within 1 mm (Fig. 4). WASB [9] and TrackNetV2 [15], run through the same pipeline, deviate from their lines by 2.4 and 5.7 mm root mean square, respectively.

5 The Agentic Workflow

An LLM-based coding agent is a large language model (LLM) that, besides producing text, runs programs, reads and writes files and starts further agents. The software that gives the model these abilities is called an agent harness, in this work Claude Code from the company Anthropic. The most capable model, the orchestrator in Fig. 5, breaks the goal set by the authors into subtasks and delegates them to subagents running a smaller model. In this work Claude Fable 5 and 5.1 served as the orchestrator and Claude Opus 4.8 and 5 as the subagents. A subagent works on one question, reports its result back and writes it into the knowledge base, a folder of plain text files, so that later agents begin with the findings of earlier ones. The knowledge base is kept deliberately compact, since a model holds only a limited amount of text at once, its working memory, and every page read costs money. Beside the knowledge base lies the literature, primary sources collected from the web, against which an agent can check a statement rather than rely on what it recalls from training. An agent that needs a source not yet in the collection searches for it and adds it to the literature.

Set-ups such as the reference rig give the agents access to real experiments, in which their hypotheses are confirmed or refuted by measurement (Section 3). This loop of hypothesis and measurement is the core of the method proposed here. The agents render their findings

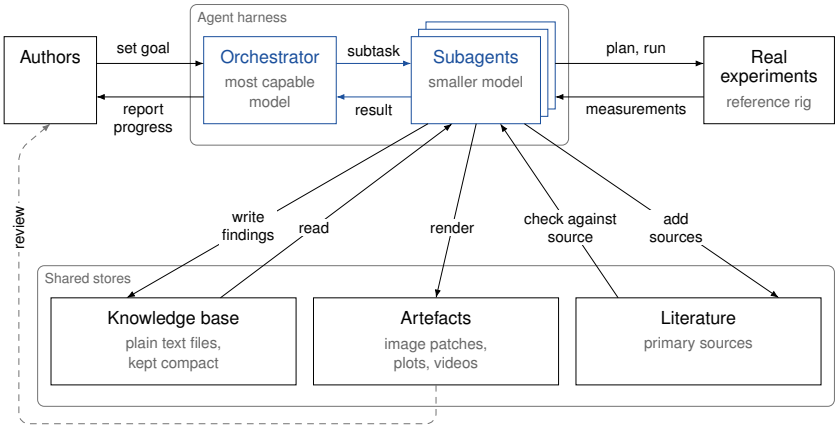


Figure 5: Organization of the work. Boxes denote the authors, the two levels of agents, the shared stores of the project and the real experiments, arrows the flow of goals, subtasks, results, findings and measurements. The dashed arrow marks the authors' review of the artefacts.

as patches, plots and videos, and the orchestrator reports progress at milestones, so the authors can follow the intermediate results. Such a loop runs for hours, at times for days, without intervention. Rendering, training and evaluation ran on one laptop, a MacBook Pro with an M1 Max chip and 64 GB of memory.

6 Conclusion, Limits and Outlook

The three steps set out in the introduction lead to a refiner that runs behind an existing coarse detector and evaluates a patch in 0.2 ms on the neural engine of an iPhone 16 Pro. It was trained on synthetic data alone and tested against a caliper-measured rig, where the revisions of the training data by the agentic workflow brought the median error from about 2 px to below 0.3 px. On real rallies filmed with two smartphones it cut the scatter of the triangulated ball positions from 5 to 1 mm.

What this paper does not show is absolute accuracy. In the hall, the agreement of two cameras measures precision only, the flight of

the serve was approximated by a straight line with its slight curvature neglected, and an error common to both cameras or to the refiner itself remains invisible. On the rig, only selected sources of error were measured and corrected, at one rotational speed and against a dark background.

An absolute reference is the subject of follow-up work. High-speed cameras around the table will read the spin off the ball's logo from several perspectives at 1 000 frames per second. These cameras give a measurement independent of the ball centres and the refiner, against which the procedure described here can be validated. The agentic workflow presented here is not limited to the specific localization problem. It applies wherever a neural network is trained on synthetic data from a physical model that has to be revised against real experiments.

Use of AI

Beyond the agentic workflow described in Section 5, AI was used to improve the language of this paper, phrasing it more clearly and concisely.

References

1. T. Gossard, J. Tebbe, A. Ziegler, and A. Zell, "SpinDOE: A ball spin estimation method for table tennis robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 5744–5750.
2. J. Tebbe, L. Klamt, Y. Gao, and A. Zell, "Spin detection in robotic table tennis," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9694–9700.
3. S. Tamaki, S. Yamagata, and S. Hashizume, "Spin measurement system for table tennis balls based on asynchronous non-high-speed cameras," *International Journal of Computer Science in Sport*, vol. 23, no. 1, pp. 37–53, 2024.
4. Spinsight ESN Digital GmbH. (2026) Spinsight. Accessed 2026-09-12. [Online]. Available: https://spinsight.com/
5. T. Gossard, A. Ziegler, and A. Zell, "TT3D: Table tennis 3D reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*, 2025, pp. 5821–5831.

6. D. Kienzle, R. Schön, R. Lienhart, and S. Satoh, "Towards ball spin and trajectory analysis in table tennis broadcast videos via physically grounded synthetic-to-real transfer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025, pp. 5832–5841.
7. D. Kienzle, K. Ludwig, J. Lorenz, S. Satoh, and R. Lienhart, "Uplifting table tennis: A robust, real-world application for 3D trajectory and spin estimation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026, pp. 7440–7449.
8. Y.-C. Huang, I.-N. Liao, C.-H. Chen, T.-U. Ik, and W.-C. Peng, "TrackNet: A deep learning network for tracking high-speed and tiny objects in sports applications," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8.
9. S. Tarashima, M. A. Haq, Y. Wang, and N. Tagawa, "Widely applicable strong baseline for sports ball detection and tracking," in *34th British Machine Vision Conference (BMVC)*. BMVA, 2023, paper 0310. [Online]. Available: https://papers.bmvc2023.org/0310.pdf
10. T. Gossard, F. Radovic, A. Ziegler, and A. Zell, "BlurBall: Joint ball and motion blur estimation for table tennis ball tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*, 2026, pp. 9930–9939.
11. L. Dong, Y. Yang, H. Wu, W. Wang, Y. Hou, Z. Zhong, and X. Sun, "RacketVision: A multiple racket sports benchmark for unified ball and racket analysis," in *AAAI Conference on Artificial Intelligence (AAAI-26)*, vol. 40, no. 5, 2026, pp. 3632–3640.
12. International Table Tennis Federation. (2025) The laws of table tennis, ITTF statutes 2025, chapter 2. pp. 38–44, accessed 2026-09-08. [Online]. Available: https://documents.ittf.sport/sites/default/files/public/2025-02/2025_ITTF_Statutes_clean_version.pdf
13. R. Schmidts. (2024) Maximal verbessert? iPhone 16 Pro Max – Sensortest – Rolling Shutter und Dynamik. slashCAM, 18 October 2024, accessed 2026-09-05. [Online]. Available: https://www.slashcam.de/artikel/Test/iPhone-16-Pro-Max---Sensortest---Rolling-Shutter-und-Dynamik.html
14. S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.
15. N.-E. Sun, Y.-C. Lin, S.-P. Chuang, T.-H. Hsu, D.-R. Yu, H.-Y. Chung, and T.-U. Ik, "TrackNetV2: Efficient shuttlecock tracking network," in *International Conference on Pervasive Artificial Intelligence (ICPAI)*, 2020, pp. 86–91.

Synthetic training data for YOLO-based optical 3D inspection of fork welds in manufacturing

Jonathan Zender[13], Stefan Maier[2], and Michael Layh[1]

[1] Hochschule Kempten, Institut für Maschinelles Sehen,
Bahnhofstr. 61, 87435 Kempten, Germany
[2] Swoboda Wiggensbach KG,
Max-Swoboda-Straße 1, 87487 Wiggensbach, Germany
[3] Universität Stuttgart, Institut für Technische Optik,
Pfaffenwaldring 9, 70569 Stuttgart, Germany

Abstract Modern manufacturing faces increasing challenges due to labor shortages, rising costs, and stricter quality requirements. In fork weld inspection, high geometric variability limits rule-based methods and makes manual inspection costly and inconsistent. AI-based inspection offers a solution, but its adoption is constrained by the lack of annotated data, especially before production start.This work proposes a synthetic data generation framework based on digital twin principles. Parametric 3D weld meshes are generated from boundary patterns representing valid geometric variations of nominal and defective welds. The approach enables high-variance geometry generation, multi-sensor simulation through physically based rendering, heightmap extraction, and automatic annotation without requiring real production data. Synthetic heightmaps achieve high similarity to real measurements, with a maximum FID of 3.12. The results demonstrate the potential of digital twin-based synthetic data generation for scalable and data-efficient AI-driven weld inspection.

Keywords 3D inspection, synthetic data generation, digital twin, YOLO, weld defect detection

1 Introduction

Industrial manufacturing is undergoing significant transformation due to workforce shortages, rising costs, and increasing quality requirements. These challenges are particularly evident in quality assurance, where defects exhibit large variations in shape, size, texture, and appearance, limiting the applicability of rule-based image processing methods [1–3]. Consequently, visual inspection has traditionally relied on human operators, which is flexible but costly, time-consuming, and prone to inconsistency [4, 5]. Deep Learning-based inspection systems offer a promising alternative; however, their deployment is often hindered by limited training data availability, expensive annotation processes, and the lack of real samples before production start [6, 7]. This work proposes a method for generating synthetic heightmap data for training machine learning models in visual inspection tasks. The approach generates 3D meshes from boundary patterns to simulate different defect types. Using fork welds as a case study, OK and NOK geometries are modeled and rendered to simulate various sensor configurations. The generated data enable automatic labeling and the creation of training-ready datasets for defect inspection.

2 Related Work

Synthetic data generation has shown potential for industrial defect inspection. Fulir et al. combine procedural surface modeling and physically based rendering (PBR) to generate realistic defect images, achieving an F1 score of 19.4% with a segmentation model trained only on synthetic data [8]. Zender et al. [9] extend this concept to laser weld inspection by generating PBR-based synthetic weld images with varied geometry, materials, surface noise, and illumination. A YOLO11 classifier trained exclusively on synthetic data achieved an F2 score of 97.7% , with topography, surface noise, and illumination identified as critical factors. Other studies focus on real-data-based weld inspection. Liu et al. [10] improve resistance spot weld detection using an enhanced Faster R-CNN, achieving 90% mAP on 1000 annotated images after augmentation. Feng et al. [11] address class imbalance in laser weld inspection with YOLO-Weld, achieving 84.8% mAP@50 and an

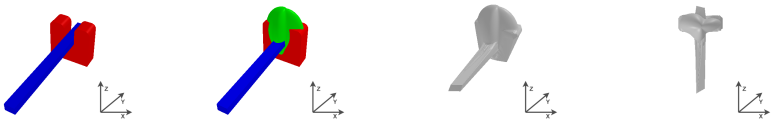
F1 score of 80.9%. Zhang et al. Overall, synthetic data can improve AI-based inspection, especially for rare defects. However, its application to fork weld inspection remains unexplored, motivating physics-based and geometry-aware synthetic data generation approaches to overcome limited defect datasets.

3 Fork Weld Generation

The fork welding process involves the insertion of the sensor leg into the fork, followed by the welding of the fork spikes to the sensor leg, see Figure 1(a). To simulate this process, a weld mesh is generated based on the boundary pattern described in the following sections. The boundary pattern defines the geometry of the weld spot, which can represent either a conforming (OK) or nonconforming (NOK) weld depending on the specified parameters. The generated weld mesh is then combined with the CAD models of the fork and the sensor leg, as shown in Figure 1(b). Finally, the resulting mesh is smoothed using Laplacian smoothing to ensure a geometrically consistent transition between the generated weld mesh and the meshes of the fork and sensor leg. See Figure 1(c)-1(d).
The welding procedure can result in common defects, including incomplete welding on a single side, tearing of the sensor leg due to excessive forces during welding, and undesired adhesion caused by melted excess fork material bonding to the sensor leg. Those typical defects are simulated by adding different mesh deformations to the generated weld mesh, which are described in more detail in section Error Generation.

4 Mesh Generation based on Boundary Patterns

A generic method for generating 3D meshes based on boundary patterns is presented, enabling the creation of various weld geometries through appropriate pattern definitions. In this work, the approach is applied to generate 3D meshes of fork welds; however, due to its boundary pattern–based formulation, it can be readily transferred to other use cases where object geometry can be described in a similar manner.



(a) 3D fork and leg without weld **(b)** 3D fork and leg with weld before smoothing **(c)** 3D weld after smoothing. **(d)** Top view of 3D weld after smoothing.

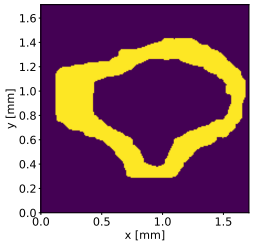
Figure 1:

4.1 Boundary Pattern

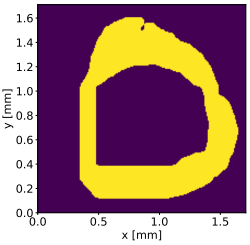
The boundary patterns defines the minimum and maximum extents of the generated mesh in x and y directions. It is required to generate a closed object and serves as the basis for 3D mesh generation. Each boundary pattern is defined in a two-dimensional coordinate system, where the x-axis represents the object width and the y-axis represents the object length. Boundary patterns can either be created manually by drawing or extracted from real-world scans of the target objects. The boundary patterns are subsequently used to generate a 3D mesh, in which the x–y geometry follows the boundary pattern, while the z-coordinates are assigned according to the radius of a sphere, which must be defined. The metric scale of the resulting meshes is defined by a multiplicative factor with millimeters per pixel as unit. Figure 2(a) shows the boundary pattern of an OK weld, whereas Figure 2(b) shows the boundary pattern of a half-welded defect.

4.2 Mesh Generation

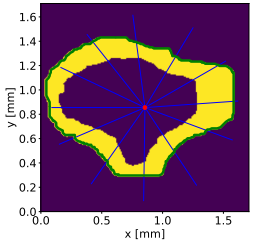
To generate a 3D mesh from the boundary pattern, the following steps are performed. First, the boundary pattern is used to determine the maximum radial distance from the center of the pattern to its outer hull, illustrated by the green line in Fig. 2(c). The arc length of the outer hull is then approximated by summing the Euclidean distances between adjacent intersection points along the boundary. The total arc



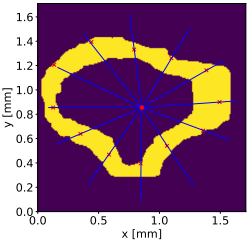
(a) Boundary pattern for an OK weld.



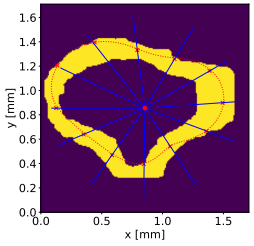
(b) Boundary pattern for a half-welded defect.



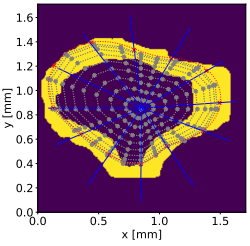
(c) Outer hull intersections (green), boundary center (red), and sampled angles (blue).



(d) Sampled angle rays (blue) and selected boundary points (red).



(e) Fitted hull (red) generated from the selected boundary points.



(f) Mesh vertices generated according to the selected mesh resolution.

Figure 2: Generation of fork weld boundary patterns and meshes. The first row shows the boundary patterns for an OK weld and a half-welded defect. The second and third rows illustrate the sampling of boundary points, hull generation, and mesh discretization.

length is then divided by the desired number of angular samples to determine the target spacing between neighboring intersection points. Based on the target spacing, a subset of angular rays is selected. The resulting selected angles are depicted as blue rays in Figs. 2(c)–2(f). Next, the selected angles are projected onto the boundary pattern, and for each angular ray a point within the boundary region is randomly selected. Example sampled points are shown as red crosses in Fig. 2(d). Finally, a cubic spline fitting function is applied to connect the selected points and form a continuous contour representing the xy- geometry of the mesh, see Fig. 2(e). Based on this contour, a 3D mesh is generated. Example vertex positions of the resulting mesh are shown in Fig. 2(f), where they are connected by contour lines.

5 Heightmap Acquisition

Due to the optical characteristics of the weld surface and its surrounding background, the geometry must be evaluated in three dimensions. This is achieved using a structured-light system, in which a line pattern is projected onto the weld by a stripe-light projector and captured by a camera. For this study, an XR-HT15M system from Keyence is employed, covering a spatial area of 12.5mm ×12.5mm with a spatial resolution of 9.5μ m. The deformation of the projected line pattern encodes the surface topology, enabling precise reconstruction of the weld geometry. To compare the generated meshes with real-world data, a dataset of real heightmaps of fork welds is acquired using the structured-light system. The dataset includes various weld conditions, such as OK welds, see Figure 3(a), half welded, adhesion on weld, adhesion on leg, ripped leg and weld dot on leg. The amount of samples for each class is shown in Table 1. Which shows the difficulty in obtaining a sufficient amount of real-world samples for training machine learning models, especially for rare defect classes such as ripped leg or weld dot on leg. This highlights the need for synthetic data generation to augment the dataset and enable effective training of machine learning models for weld defect detection. In section Results the similarity between the synthetic generated heightmaps and the real heightmaps is investigated.

Table 1: Amount of real heightmap samples for each class of the dataset.

Class	N
Ripped Leg	1
Adhesion on Leg	3
Half Welded	7
Adhesion on Weld	10
OK	146

5.1 Synthetic Heightmap Acquisition

To generate synthetic heightmaps of the welds, Blender with the Cycles renderer is used. In Blender, heightmaps can be rendered by configuring the material's BSDF parameter to correspond to the z-coordinate of the mesh. In this configuration, the rendered image represents the heightmap of the mesh, as the z-value at each point on the surface is recorded when the camera rays intersect with the mesh. This approach directly encodes surface elevation into pixel values, producing a precise representation of the weld topo

6 Error Generation

In this type of weld the most common errors to occure are

- Ripped leg
- Adhesion on leg
- Half welded
- Adhesion on weld

In order to simulate these errors, error generators are introduced. Where only one error (half welded) depends on a new boundary pattern and for the other errors it is required to generate new meshes based on mesh deformations like Perlin noise or simple geometric deformations.

6.1 Ripped leg

A ripped leg is simulated by removing part of the leg geometry below the weld, resulting in a correctly welded joint with a damaged leg. Real and synthetic examples are shown in Figure 3(c)-3(d).

6.2 Adhesion on leg

Adhesion on the leg is simulated by adding a small Perlin noise-deformed mesh to the leg. Examples are shown in Figure 3(e)-3(f).

6.3 Half Welded

A half-welded defect is simulated using a boundary pattern that defines only one weld side. A small Perlin noise-deformed mesh is added to reproduce the remaining adhesion on the unwelded side. Examples are shown in Figure 3(g)-3(h).

6.4 Adhesion on weld

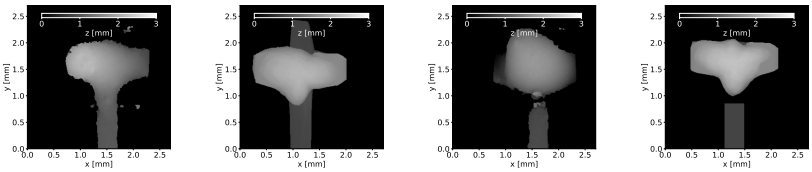
Adhesion on the weld is simulated by placing a Perlin noise-deformed spherical or elliptical mesh onto the weld surface, reproducing both dot- and spike-shaped adhesions. Examples are shown in Figure 3(i)-3(j).

7 Results

Due to the limited number of real heightmaps (Table 1), synthetic and real heightmaps are compared using SSIM [12], LPIPS [13], and FID [14].

7.1 Similarity scores

SSIM measures structural similarity (higher is better), LPIPS perceptual similarity using a pretrained VGG16 [15] (lower is better), and FID the distance between image distributions (lower is better). SSIM and LPIPS are averaged over 100 synthetic-real image pairs per class, while FID is computed between the same synthetic samples and the complete

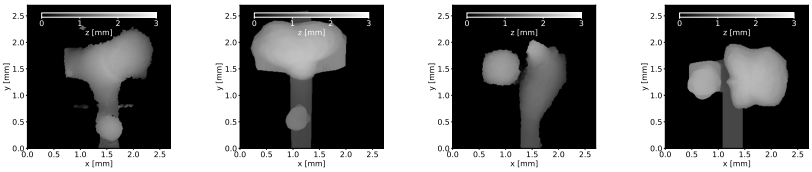


(a) Real fork weld without error

(b) Synthetic fork weld without error

(c) Real fork weld with ripped leg

(d) Synthetic fork weld with ripped leg

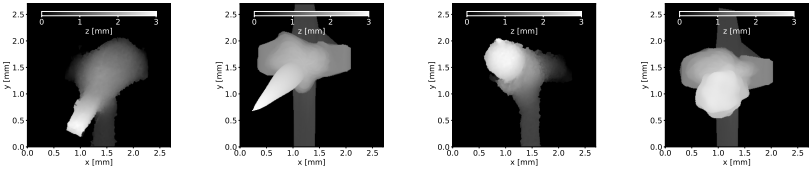


(e) Real fork weld with adhesion on leg

(f) Synthetic fork weld with adhesion on leg

(g) Real fork weld with half-welded error

(h) Synthetic fork weld with half-welded error



(i) Real fork weld with spike adhesion

(j) Synthetic fork weld with spike adhesion

(k) Real fork weld with dot adhesion

(l) Synthetic fork weld with dot adhesion

Figure 3: Comparison of real and synthetic fork weld heightmaps for the generated weld conditions. Each real weld (left) is shown next to its corresponding synthetic counterpart (right).

real dataset. For Line 1, the real OK class is randomly split to obtain balanced datasets. Results are summarized in Table 2.

Table 2: Quantitative comparison between real and synthetic images.

Class	**LPIPS**	**SSIM**	**FID**
Value	$mean \pm std$	$mean \pm std$	*score*
Real Ok 2 Real Ok	0.22 ± 0.04	0.82 ± 0.04	0.05
Adhesion on Weld	0.36 ± 0.03	0.69 ± 0.05	3.12
Half Welded	0.34 ± 0.03	0.73 ± 0.04	1.87
OK	0.33 ± 0.03	0.75 ± 0.06	1.48
Ripped Leg	0.33 ± 0.03	0.72 ± 0.05	–
Weld Dot on Leg	0.34 ± 0.03	0.73 ± 0.05	1.13

In table 2 calculated similarity scores are shown. The image based similarity scores indicate that the generated synthetic heightmaps of the welds exhibit a high degree of similarity to the real heightmaps, as reflected in the relatively low LPIPS values and high SSIM values. The FID values further support this conclusion, with values below 5 indicating a close match between the distributions of real and synthetic images. These results suggest that the proposed method for generating synthetic heightmaps based on 3D meshes derived from boundary patterns is effective in producing realistic representations of welds that closely resemble real-world data.
Furthermore, a YOLO classifier trained exclusively on synthetic height maps achieved a precision of 0.94 on real-world binary OK/NOK classification, demonstrating the effectiveness of the proposed sim-to-real approach.

8 Discussion and Outlook

The proposed method currently defines the weld geometry using a boundary pattern in the x–y plane, resulting in a spherical approximation in the z-direction. While sufficient for fork welds, this limits applicability to more complex geometries. Future extensions could incorporate an additional boundary pattern in the y–z plane to enable more flexible and realistic 3D shape modeling. Further improvements could address the angular sampling strategy by introducing geomet-

ric constraints between neighboring points to avoid abrupt transitions and improve mesh regularity. Additionally, the current cubic spline fitting does not explicitly enforce physical constraints such as curvature or slope continuity. A constraint-based fitting approach could therefore further improve geometric fidelity. Although similarity metrics demonstrate strong agreement between synthetic and real heightmaps, the generalization of models trained solely on synthetic data to real measurement data remains to be validated. Future work should evaluate cross-domain performance to confirm the applicability of the proposed synthetic data generation pipeline for industrial inspection.

9 Conclusion

This study presents a method for generating synthetic weld heightmaps based on 3D meshes derived from boundary patterns. Weld defects are simulated through mesh deformations and the addition of error geometries, while variation is increased through new boundary patterns. The generated heightmaps show high similarity to real weld measurements, achieving LPIPS values of around 0.34, SSIM values of 0.74, and a maximum FID of 3.12. The proposed approach enables the generation of large synthetic datasets for training and evaluating machine learning models for weld defect detection and classification. A YOLO classifier trained exclusively on the synthetic heightmaps achieved a precision of 0.94 on real binary OK/NOK weld classification, demonstrating the capability of the generated data to support sim-to-real transfer for industrial inspection tasks.

References

1. J. Yang, S. Li, Z. Wang, H. Dong, J. Wang, and S. Tang, "Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges," *Materials*, vol. 13, no. 24, p. 5755, 2020.
2. F. Kähler, O. Schmedemann, and T. Schüppstuhl, "Anomaly detection for industrial surface inspection: Application in maintenance of aircraft components," *Procedia CIRP*, vol. 107, pp. 246–251, 2022.
3. D. Schoepflin, D. Holst, M. Gomse, and T. Schüppstuhl, "Synthetic training

data generation for visual object identification on load carriers," in *Procedia CIRP*, vol. 104, 2021, pp. 1257–1262.

4. D. Caballero-Ramirez, Y. Baez-Lopez, J. Limon-Romero, G. Tortorella, and D. Tlapa, "An assessment of human inspection and deep learning for defect identification in floral wreaths," *Horticulturae*, vol. 9, p. 1213, 2023.
5. J. E. See, "Visual inspection reliability for precision manufactured parts," *Human Factors*, vol. 57, pp. 1427–1442, 2015.
6. N. Hütten, M. Alves Gomes, F. Hölken, K. Andricevic, R. Meyes, and T. Meisen, "Deep learning for automated visual inspection in manufacturing and maintenance: A survey of open-access papers," *Applied System Innovation*, vol. 7, p. 11, 2024.
7. J. M. Rožanec, E. Trajkova, P. Dam, B. Fortuna, and D. Mladenić, "Streaming machine learning and online active learning for automated visual inspection," *IFAC-PapersOnLine*, vol. 55, pp. 277–282, 2022.
8. J. Fulir, L. Bosnar, H. Hagen, and P. Gospodnetić, "Synthetic data for defect segmentation on complex metal surfaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023, pp. 4424–4434.
9. J. Zender, S. Maier, A. Herkommer, and M. Layh, "Synthetic data generation for ai-based quality inspection of laser welds in lithium-ion batteries," *Sensors*, vol. 25, no. 23, p. 7301, 2025.
10. W. Liu, J. Hu, and J. Qi, "Resistance spot welding defect detection based on visual inspection: Improved faster r-cnn model," *Machines*, vol. 13, no. 1, p. 33, 2025.
11. J. Feng, J. Wang, X. Zhao, Z. Liu, and Y. Ding, "Laser weld spot detection based on yolo-weld," *Scientific Reports*, vol. 14, p. 29403, 2024.
12. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
13. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
14. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017.
15. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

TorchGCvM: Fast Cramér-von-Mises distance for PyTorch

Lukas Dippon[1] and Johannes Meyer[1]

[1] Fraunhofer-Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany

Abstract The Cramér-von-Mises distance (CvMD) is a powerful metric for comparing probability distributions. Its generalized version applicable to multivariate and discrete distributions has proven useful in the context of image processing and computer vision. While its computational complexity for images of size $N \times N$ has recently been reduced from $O(N^5)$ to $O(N^3)$ by the fastGCvM algorithm, the fact that it relies on direct element-wise operations over all pixels made it impractical for use within modern deep learning frameworks like PyTorch. To allow for use of the CvMD within PyTorch, we present TorchGCvM, an efficient implementation of the CvMD, leveraging PyTorch's tensor processing and automatic differentiation. We provide both pseudocode and a reference implementation, as well as an evaluation of its performance compared to a naive implementation, demonstrating significant speedups while maintaining numerical accuracy.

Keywords Loss function, machine learning, PyTorch, Cramér-von-Mises distance, CvMD, integral images, tensor operations, optimization, automatic differentiation

1 Introduction

In any gradient-based optimization problem, the choice of a suitable loss function is crucial for the convergence and performance of the optimization process. In the context of comparing probability distributions, the Cramér-von-Mises

distance (CvMD) has proven to be a powerful metric [1]. The generalization of the CvMD to multidimensional and discrete distributions has allowed for it's application in the context of image processing and computer vision [2]. Since the CvMD computes differences and maintains gradients over increasingly larger neighborhoods, it is more robust to defocus blur and noise. This makes it an ideal candidate for optimizing images within optical simulation software [3].

To be used effectively in the context of deep learning, the CvMD needs to be implemented in a way that allows for efficient computation and automatic differentiation. While frameworks like PyTorch [4] provide efficient tensor operations and automatic differentiation, the naive implementation of the CvMD is not suitable for use within such frameworks due to its reliance on direct element-wise operations over all pixels. In the following, we first introduce the generalized discrete CvMD distance and its efficient computation using integral images, as proposed by Meyer et al. [5]. We then present the TorchGCvM algorithm, which reformulates the CvMD in a way that allows for efficient computation using tensor operations in PyTorch, while also allowing for automatic differentiation. Finally, we provide a reference implementation of the TorchGCvM algorithm and evaluate its performance compared to a naive implementation, demonstrating significant speedups while maintaining numerical accuracy.

2 Generalized Cramér-von-Mises distance

We consider the special case of comparing two discrete distributions $\mathbf{F}, \mathbf{H}$ over a 2D domain – as is common for image processing – represented as matrices of size $M \times N$. The generalized Cramér-von-Mises distance in this case is defined as the sum of squared differences between local cumulative distribution functions (LCDs) computed over square neighborhoods of increasing size around each pixel [6]:

$$d_{\mathrm{CvM}}(\mathbf{F}, \mathbf{H}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{b=0}^{B} (\mathrm{LCD}(\mathbf{F}, m, n, b) - \mathrm{LCD}(\mathbf{H}, m, n, b))^2. \tag{1}$$

Here, generally $B = \max(M, N) - 1$, and LCD describes the local cumulative distribution function of a given pixel (m, n) and neighborhood size b as

$$\mathrm{LCD}(\mathbf{A}, m, n, b) = \sum_{x=m_b^-}^{m_b^+} \sum_{y=n_b^-}^{n_b^+} \mathbf{A}_{x,y} \tag{2}$$

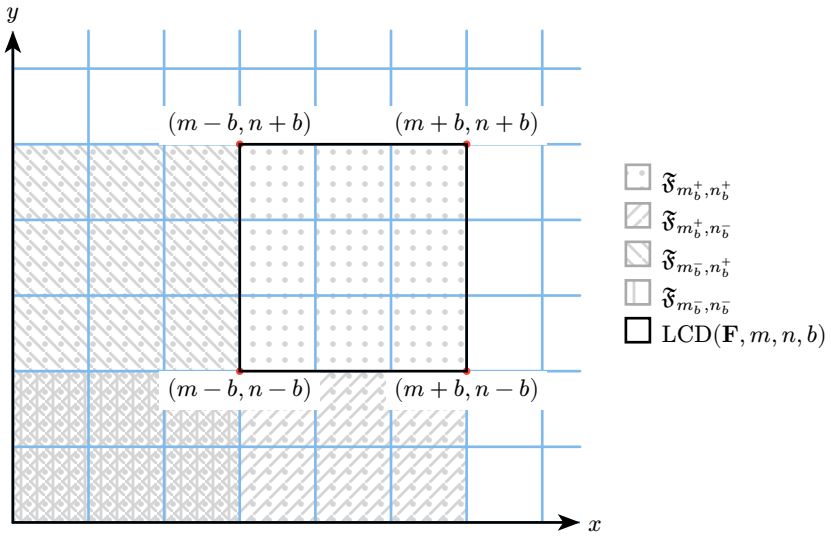


Figure 1: Region summed over for LCD with $m = 5, n = 4, b = 1$ (black outline) and corresponding areas of the integral image (gray patterns).

where $m_b^- := \max(1, m-b), m_b^+ := \min(M+1, m+b+1)$ and equivalent for n to account for boundary conditions. The naive implementation of this formula requires 5 nested loops, leading to a computational complexity of $O(N^5)$ for images of size $N \times N$, making it impractical for use with images of reasonable resolution.

To reduce complexity to $O(N^3)$, the fastGCvM algorithm introduced by Meyer et al. [5] relies on the use of integral images $\mathfrak{F}, \mathfrak{H} \in \mathbb{R}^{M+1 \times N+1}$ defined as

$$\mathfrak{F}_{i,j} = \begin{cases} 0 & \text{if } i = 1 \text{ or } j = 1 \\ \sum_{x=1}^{i-1} \sum_{y=1}^{j-1} \mathbf{F}_{x,y} & \text{else} \end{cases} \qquad \text{equiv. for } \mathfrak{H} \tag{3}$$

to compute the LCDs efficiently. Precomputing the integral images allows for constant time computation of the LCDs

$$\mathrm{LCD}(\mathbf{F}, m, n, b) = \mathfrak{F}_{m_b^+, n_b^+} - \mathfrak{F}_{m_b^-, n_b^+} - \mathfrak{F}_{m_b^+, n_b^-} + \mathfrak{F}_{m_b^-, n_b^-}. \tag{4}$$

Figure 1 illustrates the areas of the integral image that are summed and subtracted to compute the LCD for a given pixel and neighborhood size.

3 TorchGCvM

By introducing the difference of integral images $\mathfrak{D} := \mathfrak{F} - \mathfrak{H}$, the CvMD can be expressed as

$$d_{\mathrm{CvM}}(\mathbf{F}, \mathbf{H}) = \sum_{b=0}^{B} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\mathfrak{D}_{m_b^+, n_b^+} - \mathfrak{D}_{m_b^-, n_b^+} - \mathfrak{D}_{m_b^+, n_b^-} + \mathfrak{D}_{m_b^-, n_b^-} \right)^2. \quad (5)$$

Since our goal is to leverage PyTorch's tensor operations, the order of summations can be rearranged to where the summation over b is performed last, allowing for efficient computation of the inner sums over m and n using tensor operations. However, since the edge-cases require non-trivial indexing of the integral images, we introduce intermediate matrices to store the shifted integral image values for each neighborhood size b and each of the four corners of the neighborhood

$$\mathfrak{L}^{b,++} \in \mathbb{R}^{M+1 \times N+1} \quad \text{with} \quad \mathfrak{L}^{b,++}_{m,n} = \mathfrak{D}_{m_b^+, n_b^+}. \quad (6)$$

This can be efficiently computed using PyTorch's advanced indexing and broadcasting capabilities, outlined in the "shift_fill" procedure in Figure 3.

This leaves the CvMD to be computed as the sum over L_1 norms of the element-wise squared differences of the shifted integral images for each neighborhood size b:

$$d_{\mathrm{CvM}}(\mathbf{F}, \mathbf{H}) = \sum_{b=0}^{B} \left\| \mathbf{D}_b \circ \mathbf{D}_b \right\|_1, \quad \mathbf{D}_b = \left(\mathfrak{L}^{b,++} - \mathfrak{L}^{b,-+} - \mathfrak{L}^{b,+-} + \mathfrak{L}^{b,--} \right) \quad (7)$$

where $\circ$ refers to element-wise multiplication. This formulation allows for efficient computation of CvMD using tensor operations in PyTorch (illustrated by pseudocode in Figure 3), avoiding explicit loops over pixels and channels.

4 Results

To evaluate the performance of our implementation, we compare its runtime for input matrices of varying size against

- the naive implementation of the CvMD in C# as reported by Meyer et al. [5]
- the fastGCvM algorithm (also implemented in C#) by Meyer et al. [5]
- and an implementation of the fastGCVM in Python/PyTorch.

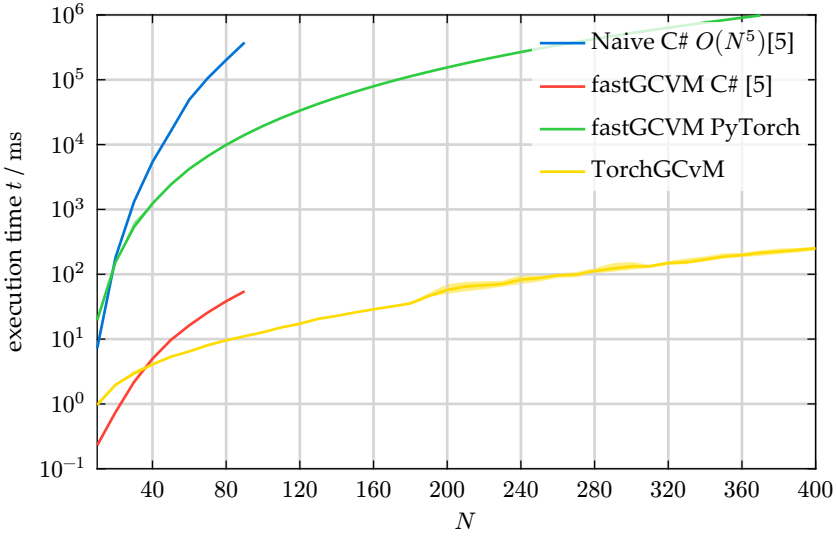


Figure 2: Execution time of different CvMD implementations for varying image sizes. The execution times for naive and fastGCvM implementations are taken from Meyer et al. [5]. Speedup by our TorchGCvM implementation more than makes up for overhead of porting to Python/PyTorch. TorchGCvM outperforms fastGCvM for larger matrices.

The results of this comparison are shown in Figure 2. We can see that porting the fastGCVM algorithm to python naively increases runtime by two orders of magnitude. At the same time, our implementation achieves a roughly threefold speedup over the naive Python implementation, significantly outperforming the native fastGCVM implementation for $N > 40$.

5 Discussion and outlook

The TorchGCvM algorithm we introduced allows for efficient computation of the CvMD by leveraging PyTorch's tensor operations and automatic differentiation capabilities. We have shown that our implementation achieves significant real-world speedups compared to both the naive implementation and FastGCvM for large input matrices, without sacrificing numerical accuracy.

However, the asymptotic complexity of the TorchGCvM algorithm remains $O(N^3)$, which may still be prohibitive for very large images. In our testing, we found that modifications such as a reduction of the number of neighborhood sizes b to only cover powers of two preserves many of the advantages of the CvMD while significantly reducing computation time to $O(N^2 \log(N))$. The implications of this modification on the performance of the CvMD in the context of image processing and computer vision tasks as well as quantitative comparisons of the CvMD to other loss metrics in machine learning applications is left to be explored in future work.

Bibliography

[1] O. C. Schrempf, D. Brunn, and U. D. Hanebeck, "Dirac Mixture Density Approximation Based on Minimization of the Weighted Cramer-von Mises Distance," in *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Sept. 2006, pp. 512–517. doi: 10.1109/MFI.2006.265624.

[2] J. Meyer, T. Längle, and J. Beyerer, "General Cramér-von Mises, a Helpful Ally for Transparent Object Inspection Using Deflection Maps?," in *Image Analysis*, P. Sharma and F. M. Bianchi, Eds., Cham: Springer International Publishing, 2017, pp. 526–537. doi: 10.1007/978-3-319-59126-1_44.

[3] L. Dippon, J. Meyer, R. Dienert, C. Kludt, and C.-W. Chen, "Optimizing Phase Modulators within Fabrication Constraints of Resin 3D Printing," in *3D Printed Optics and Additive Photonic Manufacturing V*, G. Von Freymann, A. M. Herkommer, and M. Flury, Eds., Strasbourg, France: SPIE, May 2026, p. 8. doi: 10.1117/12.3100000.

[4] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019.

[5] J. Meyer, *fastGCVM: A Fast Algorithm for the Computation of the Discrete Generalized Cramér-von Mises Distance*. Václav Skala - UNION Agency, 2017.

[6] U. D. Hanebeck and V. Klumpp, "Localized Cumulative Distributions and a Multivariate Generalization of the Cramér-von Mises Distance," in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Aug. 2008, pp. 33–39. doi: 10.1109/MFI.2008.4648104.

Appendix

Pseudocode of TorchGCvM algorithm

Algorithm 1: TorchGCvM

procedure CVM_DISTANCE($\mathbf{F}, \mathbf{H} \in \mathbb{R}^{M \times N}$)
  ▷ Initialize summed area tables
  $\mathfrak{F} \leftarrow \text{cumulative_sum}(\mathbf{F})$
  $\mathfrak{H} \leftarrow \text{cumulative_sum}(\mathbf{H})$
  $\mathfrak{D} \leftarrow \mathfrak{F} - \mathfrak{H}$
  $d \leftarrow 0$
  for $b \in \{0, ..., N\}$ **do**
    ▷ Shifted integral images for the current neighborhood size
    $\mathfrak{L}^{b,--} \leftarrow \text{SHIFT_FILL}(\mathfrak{D}, -b, -b)$
    $\mathfrak{L}^{b,-+} \leftarrow \text{SHIFT_FILL}(\mathfrak{D}, -b, +b)$
    $\mathfrak{L}^{b,+-} \leftarrow \text{SHIFT_FILL}(\mathfrak{D}, +b, -b)$
    $\mathfrak{L}^{b,++} \leftarrow \text{SHIFT_FILL}(\mathfrak{D}, +b, +b)$
    ▷ Contribution of the current neighborhood size to CvMD
    $\mathbf{D} \leftarrow \mathfrak{L}^{b,++} - \mathfrak{L}^{b,-+} - \mathfrak{L}^{b,+-} + \mathfrak{L}^{b,--}$
    $d \leftarrow d + \|\mathbf{D} \circ \mathbf{D}\|_1$
  end
  return d
end

procedure SHIFT_FILL($\mathfrak{D} \in \mathbb{R}^{M+1 \times N+1}$, $dx, dy \in \mathbb{N}$)
  ▷ Compute shifted summed area tables
  $ix \leftarrow \text{clamp}(\text{range}(1, M+1) + dx, 1, M+1)$
  $iy \leftarrow \text{clamp}(\text{range}(1, N+1) + dx, 1, N+1)$
  $\mathfrak{R} \leftarrow \text{selectColumns}(\mathfrak{D}, ix)$
  $\mathfrak{R} \leftarrow \text{selectRows}(\mathfrak{R}, iy)$
  return $\mathfrak{R}$
end

Figure 3: Pseudocode for the TorchGCvM algorithm, consisting of the main procedure "cvm_distance" and a helper procedure "shift_fill" that computes the shifted integral images. Note that indexing is 1-based to match the mathematical notation.

Structure of shifted integral images

Exemplary structure of the shifted integral images $\mathfrak{L}^{b,--}$ and $\mathfrak{L}^{b,++}$ for a given neighborhood size b and difference of integral images $\mathfrak{D}$:

$$\begin{aligned}\mathfrak{L}^{b,--} &= \left(\begin{array}{c|c} \mathbf{0}^{b\times b} & \mathbf{0}^{b\times N-b} \\ \hline \mathbf{0}^{b\times M-b} & \mathfrak{D}_{1:M-b,1:N-b} \end{array}\right) \\ \mathfrak{L}^{b,++} &= \left(\begin{array}{c|c} \mathfrak{D}_{b:M,b:N} & \mathfrak{D}_{M,b:N}\mathbf{1}^{1\times b} \\ \hline \mathbf{1}^{b}\mathfrak{D}_{b:M,N} & \mathfrak{D}_{M,N}\mathbf{1}^{b\times b} \end{array}\right)\end{aligned} \tag{8}$$

Uncertainty-guided dynamic routing for cost-efficient image classification

Edwin Rajoo, Steven Landgraf, and Markus Ulrich

Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Englerstr. 7, 76131 Karlsruhe, Germany

Abstract Deep learning networks have reached high accuracy in image classification, but often at the expense of high computational cost, since large models apply their full complexity uniformly to every input regardless of difficulty. We propose an uncertainty-guided dynamic routing framework pairing a lightweight network (MobileNetV3-Large) with a high-capacity network (EfficientNet-B3), routing each input individually based on uncertainty signals, entropy and confidence, from the lightweight model, together with its intermediate feature representation. The router is trained with a binary objective rewarding forwarding only when it corrects a mistake the small model would otherwise make, combined with a class-imbalance-aware weighting term and a regularizer discouraging unnecessary escalation. On Tiny ImageNet, the routed system reaches 78.54% accuracy, matching and, at the chosen operating point, slightly exceeding the standalone larger model (78.02%), while forwarding only around 31% of inputs to it – an estimated 57% compute reduction. An ablation study shows the two scalar uncertainty signals alone route as well as, or better than, the full signal set, with the 960-dimensional feature embedding contributing little once class imbalance is accounted for.

Keywords Dynamic routing, uncertainty estimation, adaptive inference, image classification, efficient deep learning

1 Introduction

Deep convolutional neural networks achieve high accuracy on image classification, but this typically comes at a steep computational cost:

larger models can handle more complex visual patterns, yet applying the full model to every input wastes computation when the input is simple. Lightweight architectures such as MobileNet trade capacity for speed; the trade-off shows up mainly on harder inputs, where a small model struggles once the object is partly hidden, cluttered, or easily confused with another class. Difficulty is a property of the data itself, not only an artifact of training. Easy and hard examples differ systematically across CNN architectures [1]. This observation motivates adaptive inference, where the computation applied to an input depends on the input itself rather than being fixed in advance [2].

Prior work on adaptive inference falls broadly into three groups. Early-exit and adaptive-depth networks attach internal classifiers so that easy inputs can exit before the final layer [3, 4]. Dynamic routing and conditional computation methods instead route each input along a different path through the network, using reinforcement learning, gating, or learned routing functions [5–7]. A third line of work uses the network's own confidence or predictive uncertainty to decide where to spend computation [8, 9].

All three directions, however, share a common constraint: they operate within a single network, modifying its architecture, adding exit branches, or acting on a fixed confidence threshold. None of them choose between independently trained models of different capacity using a target aligned directly with when the larger model actually helps, rather than a general notion of difficulty.

We address this gap by pairing a lightweight model, evaluated on every image, with a high-capacity model reserved for a subset of inputs. A router decides whether the lightweight prediction can be trusted or the input should be escalated to the bigger model. Routing more images to the large model raises accuracy but also cost, and vice versa. The larger model is not a strict upgrade; however the two models do not fail on the same images so a router can only exploit this complementarity by learning which specific cases benefit from escalation, not just which ones look difficult. Concretely, we ask whether such a router can be trained using only signals internal to the small classifier itself (its penultimate-layer representation, confidence, and predictive entropy) without any access to the large model at inference time. The router's training target follows directly from the complementarity argument above: rather than learning a general notion of difficulty, it is

rewarded only for forwarding an image when the large model would actually correct a mistake the small model makes, and is discouraged from forwarding images the small model already classifies correctly.

2 Related Work

Early-exit and adaptive-depth networks. The most direct approach attaches internal classifiers to a network so easy inputs can exit before the final layer. BranchyNet [3] uses side-branch classifiers at several depths; Shallow-Deep Networks [10] expose *overthinking* (cases where a correct answer was already available at a shallower layer) cutting inference cost by over 50% without losing accuracy; MSDNet [4] frames this as a test-time-budget problem, training classifiers of different costs along a multi-scale, densely-connected backbone and choosing among them adaptively. All three remain bound to their backbone, and adding exits means architecture-specific work repeated for each new one. **Dynamic routing and conditional computation.** A second family routes inputs along different paths through the network. McGill and Perona [6] train routing functions over a graph of learned transformations; Cai et al. [7] use a hypernetwork to score branch importance via Gumbel-Softmax; SkipNet [5] learns to skip residual blocks per input (30–90% less computation); BlockDrop [11] takes a similar reinforcement-learned approach (20% ImageNet speedup); Mixture-of-Experts [12] route each input to a sparse subset of expert subnetworks. All need considerably more complex training than ordinary supervised learning, and the resulting router is often tied closely to its backbone. **Uncertainty and confidence-based inference.** A third line of work uses the network's own uncertainty as its signal, but stops short of learning what to do with it. Guo et al. [8] showed modern networks are often poorly calibrated and proposed temperature scaling as a fix; Gal and Ghahramani [9] reinterpret test-time dropout as approximate Bayesian inference; Lakshminarayanan et al. [13] find a simple ensemble matches or beats elaborate Bayesian alternatives; Xia et al. [14] separate total uncertainty into model and data components via a Dirichlet-based framework.

3 Methodology

We consider inputs $x \in \mathbb{R}^{H \times W \times C}$ assigned to a class $y \in \{1, \ldots, K\}$, and design a routing system between two networks of different capacities: a lightweight CNN $f_s(x)$ and a high-capacity CNN $f_b(x)$. A routing function $r(x) \in \{s, b\}$ sends easy inputs to the lightweight network and difficult ones to the high-capacity network, giving $\hat{y} = f_{r(x)}(x)$ and minimizing cost while preserving accuracy. A router network decides which model handles each input (Fig. 1); both CNNs are trained independently and kept frozen while only the router is trained, for ease of deployment and modularity.

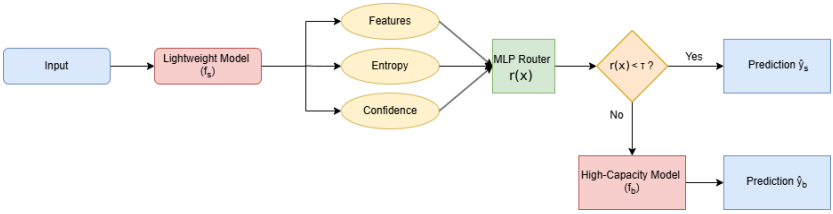


Figure 1: Overview of the proposed uncertainty-guided dynamic routing pipeline. Each input is first processed by the lightweight model f_s, which produces intermediate features, entropy, and confidence. These signals are passed to the MLP router $p(x)$, which decides whether the prediction from f_s is used directly or whether the input is forwarded to the high-capacity model f_b.

3.1 Routing Signals

The router decides not from the predicted label but from three signals describing the small model's behavior.

Intermediate features. $h(x) = f_s^{(feat)}(x)$, the 960-dimensional embedding from MobileNetV3-Large's penultimate layer after global average pooling, the same representation the model uses before its classification layer, at no extra cost. This embedding encodes how the small model internally represents the image before that representation is collapsed into a class decision, so it can carry information relevant to routing, such as the input resembling a region of feature space where the small model has historically been unreliable, even when the resulting confidence and entropy look unremarkable.

Prediction confidence. The max softmax probability of logits $z(x)$,

$$C(x) = \max_i \big(p_i(x)\big), \qquad p(x) = \text{softmax}(z(x)). \tag{1}$$

Confidence is the most direct proxy for whether the top prediction is trustworthy, but it only summarizes the probability mass assigned to a single class and says nothing about how that remaining mass is distributed across the rest. As a result, two inputs with identical confidence can differ substantially in how ambiguous the underlying decision actually is.

Predictive entropy. A measure of uncertainty across the whole output distribution,

$$H(x) = -\sum_{i=1}^{K} p_i(x) \log p_i(x). \tag{2}$$

Unlike confidence, entropy accounts for the shape of the entire distribution, so it can separate a case where one wrong class is nearly as likely as the top prediction from a case where the remaining probability mass is spread thinly across many classes. This makes it sensitive to a specific kind of failure mode, near-ties between competing classes, that confidence alone can miss.

Features provide semantic context, entropy captures distribution-wide uncertainty, and confidence reflects top-prediction strength – complementary signals for routing.

3.2 Router Network and Training

The router is a lightweight MLP: the three signals are concatenated into a 962-dimensional vector and passed through four fully connected layers (512, 256, 64, 1), with ReLU [15] after the first three and dropout [16] (0.2) after the first two to curb overfitting on the feature branch, which otherwise dominates the input by three orders of magnitude. A sigmoid function gives $r(x) \in [0, 1]$; x is routed to the large model whenever $r(x) \geq \tau$, selected empirically in Sec. 5.

Both CNNs are ImageNet-pretrained, fine-tuned independently with cross-entropy loss, then frozen. For input x with label y, let $\hat{y}_s = \arg\max f_s(x)$, $\hat{y}_b = \arg\max f_b(x)$; the router targets

$$g(x) = \mathbb{1}\big[(\hat{y}_b = y) \wedge (\hat{y}_s \neq y)\big], \tag{3}$$

i.e., $g(x) = 1$ only when the large model corrects a small-model mistake. This positive class is heavily underrepresented (about 0.5% of samples), so BCE is weighted by $w_{pos} \approx 199$ from the training-set balance:

$$\mathcal{L}_{BCE} = -\Big[w_{pos} \cdot g(x) \log r(x) + \big(1 - g(x)\big) \log \big(1 - r(x)\big)\Big]. \quad (4)$$

Two regularizers are added: a compute-aware term, and a term penalizing *small-better* cases $s(x) = \mathbb{1}[(\hat{y}_s = y) \wedge (\hat{y}_b \neq y)]$, where escalating would hurt accuracy:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda_{compute} \cdot \mathbb{E}[r(x)] + \lambda_{wrong} \cdot \mathbb{E}\big[r(x) \cdot s(x)\big], \quad (5)$$

At inference, each input passes through the lightweight CNN; the resulting signals are passed to the router, whose thresholded output decides whether to keep the lightweight prediction or escalate. We report small- and large-model accuracy, routed system accuracy, routing ratio, and router accuracy (agreement with $g(x)$).

4 Experimental Setup

Dataset. Tiny ImageNet [17]: 200 classes at native 64×64 resolution, a manageable alternative to full ImageNet [18]. Both backbones are pretrained at 224×224, so images are upsampled; training uses random cropping (4-pixel padding) and horizontal flipping, with standard ImageNet normalization.

Models. The lightweight classifier is a pretrained MobileNetV3-Large [19]; the high-capacity classifier is a pretrained EfficientNet-B3 [20]. Both use ImageNet-pretrained weights with the final layer replaced for 200 classes; the router is the MLP of Sec. 3.2.

Training. All three networks are trained separately with Adam (learning rate 10^{-4}, batch size 16), the two CNNs for 30 epochs each and the router for 60 epochs; the CNNs use cross-entropy loss and the router uses the loss of Sec. 3.2, with $\lambda_{compute} = 0.01$, $\lambda_{wrong} = 1.5$, and $w_{pos} \approx 199$ from the training-set class balance.

Evaluation metrics. We report small- and large-model accuracy, routed accuracy, routing ratio, and router accuracy (agreement with $g(x)$).

Threshold sweep, ablations, and compute cost. After router training, τ is swept from 0.10 to 0.50 in steps of 0.01, recording routed accuracy, routing ratio, and router accuracy at each value; a separate ablation study retrains the router on all seven non-empty subsets of {features, entropy, confidence}, with the CNNs held fixed. Since every input passes through the lightweight CNN before routing, total cost is approximated as $\text{Cost}_{routed} = \text{Cost}_{small} + p_{big} \cdot \text{Cost}_{big}$, where p_{big} is the share forwarded to the large model.

5 Results and Discussion

5.1 Baseline, Threshold Selection, and Case Analysis

Table 1 summarizes the accuracy of the standalone models and the full routed system at the chosen operating point ($\tau = 0.48$, see below). The routed system edges out the standalone EfficientNet-B3 model despite forwarding fewer than a third of all validation images to it. This is not simply an accuracy/efficiency trade-off: the two backbones make partially complementary errors (Table 2), and keeping the lightweight model's correct predictions instead of overriding them produces the accuracy gain alongside the compute reduction. With profiled FLOPs at 224×224 (MobileNetV3-Large: 0.22 GFLOPs; EfficientNet-B3: 1.83 GFLOPs), the routed system averages about 0.78 GFLOPs/image against 1.83 GFLOPs for uniform large-model inference – an estimated 57% reduction.

Table 1: Baseline classification performance.

Model	Accuracy
Lightweight CNN (MobileNetV3-Large)	73.89%
High-capacity CNN (EfficientNet-B3)	78.02%
Routed System ($\tau = 0.48$)	**78.54%**

Routed system accuracy is essentially flat across $\tau \in [0.10, 0.50]$ (Fig. 2): 78.72% at $\tau = 0.10$ down to 78.52% at $\tau = 0.50$, degrading gracefully rather than sharply as the workload shifts to the lightweight model. Scoring each threshold as $\text{score}(\tau) = acc_{routed}(\tau) - 0.001 \times (100 - pct_{small}(\tau))$, a light penalty for reliance on the large model, se-

lects $\tau = 0.48$: 78.54% routed system accuracy while sending only 30.83% of images to the high-capacity model, against 35.62% at the highest-accuracy threshold.

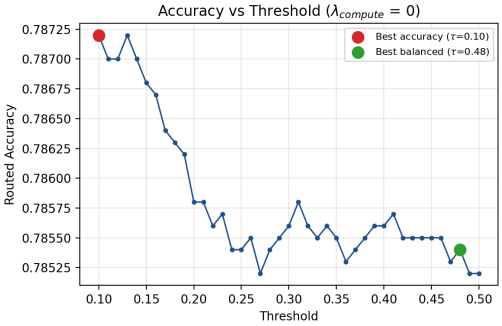


Figure 2: Routed system accuracy as a function of routing threshold. Red marker: Threshold maximizing raw accuracy. Green marker: Chosen operating point.

Breaking decisions down by outcome at $\tau = 0.10$ (Table 2): of 1105 cases where the large model would fix a small-model mistake, only 99 are missed, but of 692 cases where the small model is actually better, 502 (over 70%) are forwarded anyway – the router catches corrections far better than it avoids unnecessary escalation. This asymmetry is baked into training: $g(x)$ rewards catching every correction, while λ_{wrong} only discourages, rather than rules out, the opposite mistake. An oracle that escalates only when it fixes a mistake reaches 84.94% while sending just 11.1% of images to the large model – over six points above the learned router, showing it exploits only part of the backbones' complementary structure, leaving substantial potential for future work.

5.2 Ablation Study

Table 3 shows results for all seven signal combinations at $\tau = 0.48$. The full router is surprisingly not the best configuration: entropy-only (79.00%) and confidence-only (78.97%) both beat it (78.56%), and dropping features entirely still reaches 78.96%. Entropy is the most

Table 2: Routing case analysis on the validation set ($\tau = 0.10$).

Outcome	Count
Both models correct	6697
Large model fixes small model's error	1105
of which, missed (not routed to large model)	99
Small model correct, large model wrong	692
of which, wrongly routed to the large model	502
Both models wrong (irreducible)	1506

load-bearing signal (removing it hurts accuracy most, 76.70%); confidence tracks it closely, suggesting the most useful information already sits in the top softmax probability. Features alone reach only 74.69% while pushing 86.3% of inputs to the small model – default avoidance rather than confident routing – likely because the 960-dimensional feature vector outweighs the two-dimensional signal pair by three orders of magnitude, inviting overfitting given how rare the positive class is. Router accuracy is not a trustworthy proxy for quality either: only-features has the highest router accuracy (79.56%) but the worst routed system accuracy (74.69%).

Table 3: Ablation study on router input components at $\tau = 0.48$.

Feat.	Ent.	Conf.	Sys. Acc. (%)	Small Usage (%)	Router Acc. (%)
✓	✓	✓	78.56	69.10	74.27
–	✓	✓	78.96	59.00	67.60
✓	–	✓	76.70	80.60	78.79
✓	✓	–	78.65	67.30	72.92
✓	–	–	74.69	86.30	79.56
–	✓	–	79.00	61.50	69.48
–	–	✓	78.97	62.90	70.51

5.3 Discussion

The ablation results contextualize the earlier assumption that more signal is always better. Entropy and confidence alone are not only sufficient, they are slightly better than the full combination. Features add little by themselves and seem to introduce noise once stacked on top of the other two. With only around 0.5% of training samples are positive,

this is not a huge surprise: a 960-dimensional feature vector gives the router far more surface area to overfit spurious correlations than a two-dimensional uncertainty signal does. None of this rules out feature-based signals in general; it just means their value depends on the specific backbones, dataset size, and class balance in play, not something to add by default.

The case analysis and oracle comparison point at where the remaining gap actually is. An oracle with access to ground truth reaches 84.94% while sending just 11.1% of inputs to the large model, well past what the learned router manages, and far more selective about it. Most of that gap traces back to one specific failure mode: the router is much more careful about not missing a large-model correction than about avoiding unnecessary escalation where the small model was already right. A stronger wrong-routing penalty, or a three-way target instead of a binary one, would target this asymmetry directly.

Despite impressive results, two limitations remain. Every input pays for a full lightweight-model forward pass before any routing decision is made, so this framework cannot claim savings within the lightweight model itself – a problem that could be addressed with hierarchical or early-exit extensions. Additionally, our findings are based on limited evaluations on Tiny ImageNet only. Extending our work to other datasets with different class balances or difficulty profiles, or evaluating different model architectures, presents an obvious direction for future work.

6 Conclusion

A router built from nothing but entropy and confidence, no feature access required, already matches a substantially larger classifier's accuracy while touching it for less than a third of the time, at an estimated 57% reduction in compute. Interestingly, the oracle and case-level results show that this is not the ceiling yet: a more targeted routing objective – for instance a three-way target that separately penalizes missed corrections and unnecessary escalations, rather than a single binary target that only partly discourages the latter – could improve the overall performance and efficiency even further. Extending the framework with an early-exit mechanism within the lightweight model itself, and

testing it on datasets with a different difficulty and class-balance profile than Tiny ImageNet, are natural next steps.

References

1. I. Kishida and H. Nakayama, "Empirical study of easy and hard examples in CNN training," in *Neural Information Processing*, ser. Communications in Computer and Information Science. Springer, 2019, vol. 1142, pp. 179–188.
2. Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7436–7456, 2022.
3. S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2464–2469.
4. G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *International Conference on Learning Representations (ICLR)*, 2018.
5. X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "SkipNet: Learning dynamic routing in convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 409–424.
6. M. McGill and P. Perona, "Deciding how to decide: Dynamic routing in artificial neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 2363–2372.
7. S. Cai, Y. Shu, and W. Wang, "Dynamic routing networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3588–3597.
8. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1321–1330.
9. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 48, 2016, pp. 1050–1059.
10. Y. Kaya, S. Hong, and T. Dumitraş, "Shallow-deep networks: Understanding and mitigating network overthinking," in *Proceedings of the 36th Interna-*

tional Conference on Machine Learning (ICML), ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 3301–3310.

11. Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris, "BlockDrop: Dynamic inference paths in residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8817–8826.
12. N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations (ICLR)*, 2017.
13. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
14. F. Xia, J. Snell, and T. L. Griffiths, "Early exiting in deep neural networks via Dirichlet-based uncertainty quantification," in *NeurIPS 2024 Workshop on Fine-Tuning in Machine Learning*, 2024.
15. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.
16. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
17. Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," CS231N Course Project, Stanford University, Tech. Rep., 2015.
18. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
19. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.
20. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 6105–6114.

Evaluating multi-view anomaly detection for inspection of reusable containers

Vincent Wahyudi and Martin Storath

Institute of Digital Engineering (IDEE), Technische Hochschule Würzburg Schweinfurt (THWS),
Ignaz-Schön-Straße 11, 97421 Schweinfurt, Germany

Abstract This paper explores a multi-view anomaly detection (MVAD) model that has a base architecture consisting of a student-teacher model with a multi-view adaptive selection module with correlation (MVAS-correlation) and epipolar (MVAS-epipolar) blocks. The objective of the paper is to investigate whether background removal and multi-view feature fusion can improve the model's performance on a real industrial dataset (the SLC (Small Load Carrier) Dataset) that contains clutter in the background, which is not present in benchmark industrial anomaly detection (IAD) datasets, such as the Real-IAD dataset. The experiments are conducted on one class for both datasets. The best performance in the audiojack class is achieved by using IS-Net background removal with the baseline student-teacher model, with an AUROC of 86.99 at the image level. For the sample level, the MVAS-epipolar method without background removal performs best with an AUROC of 94.58. For the KLT_18 class, IS-Net background removal with the baseline model achieved AUROCs of 91.30 and 97.50 at the image and sample levels, respectively.

Keywords Industrial anomaly detection, epipolar constraints, background removal, quality control, reusable containers

1 Introduction

Industrial Anomaly Detection (IAD) is a crucial step in manufacturing to evaluate objects, maintain sample quality, and filter out defective

samples. A common industrial application of IAD is the evaluation of Small Load Carriers (SLCs) [1]. IAD datasets initially focused on independent single-view image benchmark datasets such as MVTec-AD [2]. As the performance of 2D IAD models on the MVTec-AD improves, more complex benchmark datasets such as Real-IAD [3] are developed. Real-IAD introduces dependencies through its data collection mechanism, as the object is recorded using the cameras from multiple viewing angles simultaneously. This creates the potential to incorporate additional information such as geometrical constraints between the cameras or semantic information between the images.

Benchmark datasets such as MVTec-AD and Real-IAD are suitable for gauging the initial performance of unsupervised IAD models, as they have controlled black or white backgrounds. However, real industrial scenarios contain background clutter, such as the partial view of conveyor belts in the SLC dataset [4]. This clutter can affect IAD performance, as it can be considered an anomalous region, especially in unsupervised IAD. Unsupervised IAD uses abundant normal SLC samples for training and the remaining defective samples for evaluation. These defective samples are limited, as they require human expertise for correct labeling.

The main contribution of this paper is the evaluation of models developed for the benchmark Real-IAD dataset in a real-world use case such as the SLC dataset. We further examine the out-of-the-box background removal methods to determine whether they improve model performances by removing background clutter. Due to computational constraints, most experiments are conducted on one class from each dataset: audiojack from Real-IAD and KLT_18 from SLC.

This paper is structured as follows: Section 2 explores existing state-of-the-art methods for multi-view IAD. Section 3 evaluates MVAD methods on the Real-IAD and SLC datasets. Section 4 evaluates different out-of-the-box background removal backbones on the Real-IAD and SLC datasets and examines their impact on the metrics. Section 5 summarizes the experimental results and discusses potential future work to further improve MVAD models.

2 Related Work

IAD models initially focused on evaluating individual 2D RGB images, such as the 2D MVTec AD [2] dataset. Common 2D-RGB IAD models can be divided into several groups based on their implementation: memory bank-based methods [5, 6], student-teacher-based methods [7, 8], reconstruction-based methods [9, 10], and distribution-based methods [11, 12]. Liu et al. [13] further elaborated on these models, most of which are available for testing in the Anomalib [14] Python library.

With the saturated performance of 2D RGB IAD models on the benchmark MVTec AD dataset, more complicated datasets have been developed to simulate realistic industrial setups. Real-IAD [3] uses a five-camera setup around the object covering the top, front, back, left, and right views. This creates a multi-view AD task that establishes semantic relationships between the views. Methods that address the multi-view task include Multi-view Adaptive Selection (MVAS) [15], a student-teacher-based method with a specialized module that maps semantic correlations between views using a correlation matrix. Multi-view Epipolar Attention Anomaly Detection (MVEAD) [16] is a PatchCore [5] memory-bank-based method that uses epipolar constraints for geometrically constrained attention instead of the data-driven semantic correlations used in MVAS. Multi-Flow [17] uses normalizing flows to fuse information across views for AD tasks.

3 Evaluating MVAD Methods in Real-IAD and SLC datasets

3.1 Real-IAD and SLC Dataset: Similarities and Differences

Table 1 describes both the Real-IAD and SLC datasets, highlighting their similarities and differences. Both datasets use a similar data acquisition method, where five cameras at different positions capture views from different angles simultaneously. Differences between the datasets include the distribution of OK and NOK samples, object position during data acquisition, and background clutter, which are common in real-world scenarios. Labels in the SLC dataset are limited

Table 1: This table describes the characteristics of the Real-IAD and SLC datasets. The main similarity is the number of images per sample, while the main differences are the background conditions, OK conditions, NOK labels, and defect characteristics.

Information/Datasets	Real-IAD [3]	SLC Dataset [4]
Total No. of images	150 K	17 K
No. of OK	99 K (66%)	14 K (82%)
No. of NOK	51 K (34%)	3 K (18%)
Total No. of classes	30	34
Defect Characteristics	one defect per image	may contain more than one defect per image
No. of images per sample	5 (top and 4 sides)	5 (top and 4 sides)
Position of object	Centered	Not always centered
Background	Clean white or black	Conveyor and clutter
OK conditions	Brand new	Samples contains wear
NOK labels	Segmentation labels	Image and sample level labels
Additional details	Different kinds of industrial objects (i.e. audio-jack, PCB, USB, etc.)	Different types of SLCs such as covers, inlays, boxes

to whether samples are defective or normal and do not explicitly define the causes of defects. In contrast, Real-IAD provides masks for segmentation tasks. Another characteristic of the SLC dataset is that objects can have more than one defect in a single view.

3.2 Evaluated Architecture: Baseline Student-teacher, Correlation Blocks (MVAS-correlation) and Epipolar Constraint Blocks (MVAS-epipolar)

The current experiment implements three different architectures: the baseline student-teacher network without correlation blocks, MVAS-correlation [15], and MVAS-epipolar. MVAS is a student-teacher network with MVAS blocks that enhance single-view features using information from the remaining views. The method consists of a pre-trained teacher encoder for multi-scale feature extraction, a student decoder that mirrors the teacher's architecture, three MVAS blocks with different dimensions, and an FPN-like structure that aggregates the multi-

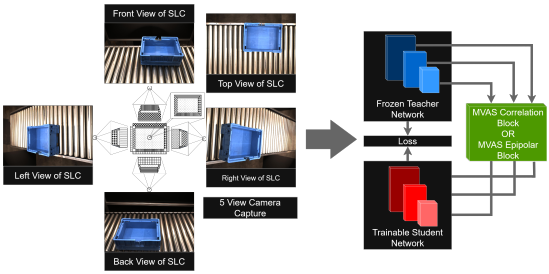


Figure 1: This is the student-teacher architecture with the MVAS-correlation or MVAS-epipolar blocks. The teacher is a frozen, non-trainable model that takes in the five views of a sample. The student model is a trainable mirror of the teacher model. The MVAS-correlation or MVAS-epipolar blocks act as feature fusion modules based on semantic correlation or epipolar geometry, respectively.

scale features.

Taking inspiration from MVEAD [16], we modified the MVAS blocks into a geometrically constrained attention module called MVAS-epipolar, based on the epipolar constraints used in multi-view camera setups. The main architecture of the student-teacher methods remains the same. The goal of the modification was to change from relying on semantic similarities, which can perform poorly with a limited training data and textureless surfaces in the SLC dataset, to more robust epipolar geometric constraints. Epipolar geometry states that when using non-degenerate stereo vision, a pixel in view a and its corresponding projection in view b must lie on the epipolar line, satisfying the epipolar constraint. The epipolar constraint requires the calculation of the fundamental matrix relating the two views, which is estimated using the eight-point algorithm [18]. Figure 1 shows the architecture used in our experiments.

3.3 Experimental Setup and Metric Evaluation

Due to computational constraints, the experiments are conducted on only one class from each dataset, which are audiojack (Real-IAD) and KLT_18 (SLC), respectively. The same training parameters are used for all experiments. All models use the ResNet34 backbone for feature extraction and are trained for 100 epochs using the AdamW optimizer.

The learning rate is 0.0125, and the weight decay is set to 10^{-4}. The metrics used for the experiments are AUROC (Area Under the Receiver Operating Characteristic Curve), AP (Average Precision), and F1-Score. However, we focus more on AUROC and AP, as they are less reliant on threshold values.

Table 2: Experimental results of the three architectures with different background removal methods evaluated using AUROC, F1-Score and AP on the audiojack class. The best-performing models are given in **bold**. Values are given as image/sample-level values. SAM background removal method consistently performs worse. There is no consistent performance improvement when changing from baseline to multi-view fusion method.

Dataset: Real-IAD (audiojack)				
Method	Rembg	AUROC	F1-Score	AP
Student-Teacher Baseline Model	None	86.20/94.36	66.98/82.31	80.94/97.17
	U^2-Net	86.65/92.70	**67.73**/80.98	81.40/95.97
	IS-Net	**86.99**/91.91	67.42/80.53	**82.59**/95.68
	SAM	68.65/85.84	52.62/75.42	61.45/90.92
MVAS (Correlation)	None	86.34/93.86	67.29/81.87	81.74/96.91
	U^2-Net	86.20/92.55	66.79/81.20	81.25/96.16
	IS-Net	86.08/92.87	66.98/80.98	80.90/96.21
	SAM	68.76/88.16	53.50/78.98	59.25/92.91
MVAS (Epipolar)	None	85.99/**94.58**	66.92/82.09	80.64/**97.31**
	U^2-Net	85.34/94.04	66.17/**82.76**	78.97/96.96
	IS-Net	85.68/94.09	67.17/81.42	79.81/96.85
	SAM	67.70/88.21	53.31/79.20	56.37/92.69

3.4 Experimental Results

Table 2 and Table 3 show the three different architecture methods trained on the audiojack (Real-IAD) and KLT_18 (SLC) datasets, respectively. The "None" in the Rembg column means that no background removal pipeline was implemented. In Table 2, the "None" rows show that there is a slight increase in performance at the image level from the baseline to MVAS-correlation and at the sample level from the baseline to MVAS-epipolar. This suggests that incorporating multi-view information using epipolar constraints may provide a slight improvement at the sample level. However, this improvement is not observed

Table 3: Experimental results of the three architectures with different background removal methods evaluated using the AUROC, F1-Score and AP on the KLT_18 class. The best performing models are given in **bold**. Values are given as image/sample-level values. Background removal methods have a bigger impact than changing the multi-view fusion methods for KLT_18 class.

Dataset: SLC (KLT_18)				
Method	Rembg	AUROC	F1-Score	AP
Student-Teacher Baseline Model	None	88.41/90.85	71.87/**73.03**	97.98/98.56
	U^2-Net	91.11/96.62	**73.03**/**73.03**	98.61/99.51
	IS-Net	**91.30**/**97.50**	72.70/**73.03**	**98.62**/**99.62**
	SAM	84.98/91.93	72.37/**73.03**	97.56/98.80
MVAS (Correlation)	None	87.62/90.08	71.87/72.20	97.87/98.40
	U^2-Net	90.91/96.87	72.86/**73.03**	98.56/99.54
	IS-Net	89.97/95.25	72.53/**73.03**	98.36/99.27
	SAM	83.70/91.62	72.20/**73.03**	97.32/98.72
MVAS (Epipolar)	None	87.82/90.11	71.87/**73.03**	97.87/98.45
	U^2-Net	90.46/96.56	72.70/**73.03**	98.46/99.47
	IS-Net	90.97/95.96	72.53/**73.03**	98.54/99.39
	SAM	83.81/91.47	72.37/**73.03**	97.32/98.67

for KLT_18, where the baseline model performs better at both the image and sample levels without background removal (see the "None" Rembg row of Table 3). To address the similar F1-Score in KLT_18, the evaluation uses the median anomaly score as threshold, which can lead to saturation when the test set is small and imbalanced. Therefore, AUROC and AP provide a more informative comparison of the models.

Figure 2 shows correct predictions of defective audiojack and KLT_18 samples without background removal. The "Final Defect Decision" shows the location of the defects, which corresponds well with the ground truth.

Figure 3 demonstrates cases where test samples are misclassified. The first two columns show OK samples from KLT_18 and audiojack that were misclassified by the model. In the KLT_18 case, the "Potential Defect Heatmap" shows that a potential defect outside the container, marked in red, which is the end of the conveyor belt. In the audiojack case, it mistakenly highlights something inside the object as defective.

The last three columns show defective KLT_18 and audiojack samples that are misclassified as OK. The location of the defects is shown in the last column "Defect Ground Truth", as boxes (self-annotated) and a binary mask. The defective KLT_18 case of misclassification has shows red markings outside the container and inside the container, but the markings inside the container are not high enough to pass the threshold, resulting in the misclassification.

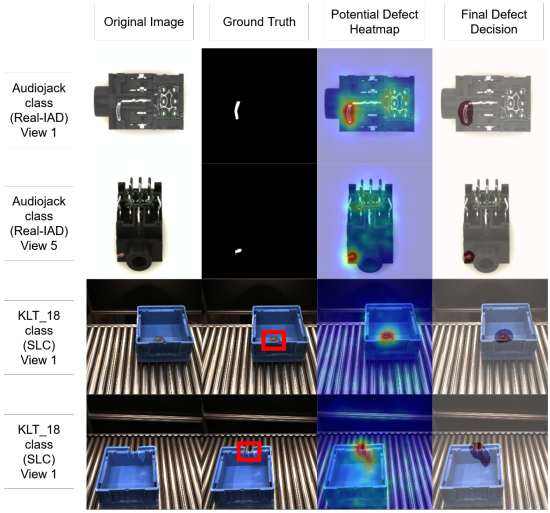


Figure 2: This is the qualitative comparison of the two datasets (KLT_18 and audiojack) on the MVAS-epipolar model. The first column is the original image, followed by the ground truth of the defect location. The ground truth for the SLC is self-annotated and represented by the red rectangle. The "Potential Defect Heatmap" column shows the heatmap of the anomaly score relative to the whole image. The red spots mean higher anomaly scores and anomaly scores that pass a particular threshold are represented in the final column as a blue mask.

4 Improving Performance using Background Removal

As there are differences in the background between the KLT_18 (SLC dataset) and audiojack (Real-IAD dataset), we implemented back-

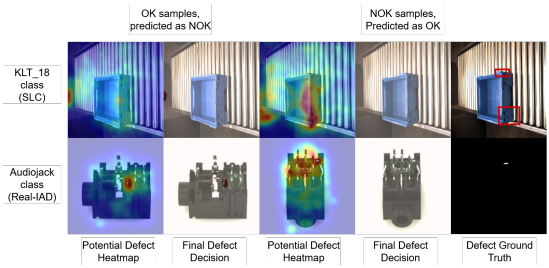


Figure 3: The first two columns represent the OK samples from KLT_18 and audiojack that are misclassified. In the KLT_18 case, the defect is falsely detected outside the container. The last three columns represent the NOK samples that are misclassified as OK. The KLT_18 case has higher anomaly scores that are present in the background of the image. Despite these scores not passing the threshold to be considered NOK, this shows that the model is focusing on both the foreground and background which affects the final decision.

ground removal methods based on U^2-Net [19], IS-Net [20], and Segment Anything Model (SAM) [21] using the Python library rembg [22]. As the audiojack images have a black/white background, the output after the background removal method is very similar to the original images, except for the SAM, which sometimes fails to differentiate between the foreground and background, causing it to crop parts of the original object. In the KLT_18 sample, the SAM also fails. Both IS-Net and U^2-Net background removal methods isolate the container properly.

Table 2 and Table 3 show the results of background removal on the three architecture models. In the audiojack case, adding background removal decreases the performance of the models at both the image and sample levels (lower AUROC and AP) for the multi-view methods, except for the baseline model. In the KLT_18 case, there is an overall increase in performance for each architecture at both image and sample level, except for the SAM background removal method in the image level. Figure 4 shows the results of background removal using U^2-Net on both audiojack and KLT_18. The previously misclassified KLT_18 OK samples in Figure 3 are now correctly classified as OK samples, as there are no highlights near the edge of the image. The defective KLT_18 samples are now correctly classified as NOK despite not de-

tecting the corner breakage. The audiojack cases remain misclassified after the background removal.

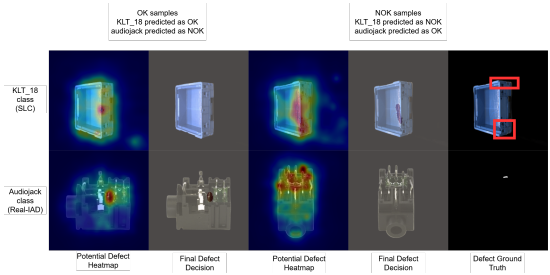


Figure 4: This is the same method with Figure 3 after U^2-Net background removal. The KLT_18 samples are now correctly classified. However, the audiojack class still experiences misclassfication.

5 Conclusion and Future Work

The change in architecture from the baseline without correlation to MVAS-correlation and MVAS-epipolar does not show consistent performance improvements for the audiojack and KLT_18 classes. Background removal is more important for the KLT_18 class, as it removes background clutter that was in some cases considered as defective regions. The best-performing combination for KLT_18 is the baseline student-teacher model with IS-Net, achieving AUROC values of 91.30 and 97.50 at the image and sample levels, respectively. As the experiments are limited to one class per dataset, further experiments on the full datasets should provide a clearer picture of the performance of the multi-view fusion architecture.

Acknowledgements

This work is supported by DIG:IT-KMU (Grant No. 20-3044-2-23), funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under the EFRE Bayern 2021-2027 programme and co-financed by the European Union.

References

1. A. Dobhan, L. Eberhardt, M. Haseneder, H. Raab, S. Rabenstein, A. Treutlein, V. Wahyudi, and M. Storath, "Digitalization in small-load-carrier management." in *ICEIS (1)*, 2025, pp. 848–855.
2. P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *International Journal of Computer Vision*, vol. 129, no. 4, p. 1038–1059, Apr. 2021.
3. C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma, "Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 22 883–22 892.
4. V. Wahyudi, C. C. Ziegler, M. Frieß, S. Schramm, C. Lang, L. Eberhardt, F. Freund, A. Dobhan, and M. Storath, "A computer vision system for recognition and defect detection for reusable containers," *Machine Vision and Applications*, vol. 36, no. 2, p. 39, Jan. 2025.
5. K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 298–14 308.
6. T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 475–489.
7. H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9737–9746.
8. G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," *arXiv preprint arXiv:2103.04257*, Oct. 2021.
9. V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8330–8339.
10. V. Zavrtanik, M. Kristan, and D. Skočaj, *DSR – A Dual Subspace Re-Projection Network for Surface Anomaly Detection*, ser. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, vol. 13691, p. 539–554.
11. J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," Nov. 2021.

12. D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *2022 IEEE/CVF winter conference on applications of computer vision (WACV)*. IEEE, 2022, pp. 1819–1828.

13. J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, p. 104–135, Feb. 2024.

14. S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A deep learning library for anomaly detection," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1706–1710.

15. H. He, J. Zhang, G. Tian, C. Wang, and L. Xie, "Learning multi-view anomaly detection with efficient adaptive selection," *IEEE Transactions on Multimedia*, 2026.

16. Y. Liu, X. Xu, S. Li, J. Liao, and X. Yang, "Multi-view industrial anomaly detection with epipolar constrained cross-view fusion," Mar. 2025.

17. M. Kruse and B. Rosenhahn, "Multi-flow: Multi-view-enriched normalizing flows for industrial anomaly detection," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2025, pp. 3933–3944.

18. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

19. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

20. X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. V. Gool, "Highly accurate dichotomous image segmentation," 2022.

21. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, p. 3992–4003.

22. D. Gatis, "rembg," 2026. [Online]. Available: https://github.com/danielgatis/rembg

Cascaded outlier rejection for object-specific monocular time-to-collision estimation

Msuega Jnr. Iorpenda, Antonio Sabotinov, and Volker Willert

Center for Robotics (CERI). Technical University of Applied Sciences,
Würzburg-Schweinfurt,
Konrad-Geiger-Strasse 2, 97421 Schweinfurt.

Abstract Monocular time-to-collision (TTC) estimation is attractive for lightweight robotic platforms but remains sensitive to tracking errors, rotational flow, and poorly conditioned focus-of-expansion (FoE) estimation. This paper presents an object-specific TTC framework with cascaded outlier rejection using forward–backward validation, epipolar filtering with rotation compensation, FoE RANSAC, and median aggregation. Perpendicular, circular, and square FoE residuals are compared across motion baselines. Gazebo experiments show that TTC accuracy depends on temporal baseline, with the preferred frame stride decreasing as approach velocity increases. Under a 5° yaw perturbation, the cascade reduces MAE by 80.8% relative to the unfiltered baseline. FoE residuals differ mainly under weak image expansion and converge under well-conditioned motion. Real DJI Tello experiments with OptiTrack ground truth achieve 0.604 s MAE, 0.211 s MedAE, and 6.566% MAPE. Robust object-specific TTC therefore requires geometric filtering and appropriate temporal-baseline selection.

Keywords Time-to-collision, optical flow, monocular vision, collision avoidance, indoor drones

1 Introduction

Autonomous micro aerial vehicles (MAVs) are increasingly used for indoor inspection, inventory, exploration, and search-and-rescue, where GNSS may be unavailable and perception must operate under tight

payload, power, and computational constraints [1]. Monocular vision is attractive in this setting because it is lightweight and inexpensive. Although metric depth is generally ambiguous from a single image [2], time-to-collision (TTC) can be inferred directly from image motion. Under translation, image features diverge from a focus of expansion (FoE), enabling collision estimation without explicit 3-D reconstruction [1, 3, 4].

Conventional TTC methods often process global or region-level image motion, allowing features from different objects, depths, or background surfaces to affect the same estimate. Object-specific TTC instead associates image motion with individual obstacles. Previous work combined object localisation, sparse optical flow, rotation compensation, FoE estimation, and monocular TTC [5]; however, tracking and motion-estimation errors can propagate through these stages and degrade the final estimate.

This work introduces a cascaded object-specific TTC framework that rejects unreliable correspondences before TTC aggregation. Shi–Tomasi features within detected bounding boxes are tracked using pyramidal Lucas–Kanade flow [3, 6]. Forward–backward validation removes inconsistent tracks [7], epipolar geometry rejects motion-inconsistent correspondences and supports rotation compensation [8], and FoE RANSAC retains tracks consistent with a common object-specific expansion centre [9]. The surviving feature-level TTC estimates are aggregated using the median. The framework compares perpendicular, circular, and square FoE residuals, with the circular and square formulations explicitly accounting for pixel uncertainty [10, 11]. Their effects on correspondence retention and TTC accuracy are evaluated alongside temporal-baseline selection and cascade ablation.

2 Cascaded Object-Specific TTC Estimation

Figure 1 summarizes the object-specific TTC pipeline. For each matched object detection, sparse correspondences are progressively filtered before feature-level TTC estimation and robust aggregation.

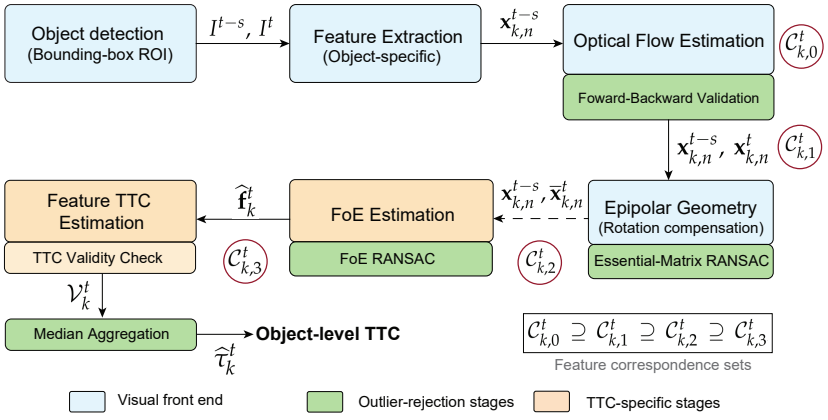


Figure 1: Cascaded object-specific monocular TTC pipeline from detected bounding-box ROI to geometric outlier rejection and robust object-level TTC estimation.

2.1 Object-Specific Correspondence Generation

Consider two monocular images I^{t-s} and I^t acquired at timestamps t_{t-s} and t_t, respectively, where s denotes the frame stride. Their temporal separation is $\Delta t_s = t_t - t_{t-s}$, which reduces to approximately s/f_{cap} for a constant capture rate f_{cap}.

An object detector provides a bounding box B_k^{t-s} for object k. Detections of the same semantic class in I^t are associated by maximum intersection-over-union (IoU), yielding the matched box B_k^t. YOLO is used in the implementation, while the subsequent geometric processing is independent of the detector architecture.

The input images are first undistorted using the calibrated camera model. Shi–Tomasi features [6] are extracted inside $\mathcal{B}_k^{t-s}$ and tracked to I^t using pyramidal Lucas–Kanade optical flow [3]. A track is retained only when forward tracking succeeds and its endpoint lies inside $\mathcal{B}_k^t$. The resulting object-specific correspondences form $\mathcal{C}_{k,0}^t$. For the nth correspondence, $\mathbf{x}_{k,n}^{t-s}$ and $\mathbf{x}_{k,n}^t$ denote the tracked feature coordinates in the undistorted image plane, with homogeneous representation $\widetilde{\mathbf{x}} = [\mathbf{x}^\top, 1]^\top$, where $\mathbf{K}$ denotes the camera intrinsic matrix. The corresponding image-flow vector is $\mathbf{v}_{k,n}^t = \mathbf{x}_{k,n}^t - \mathbf{x}_{k,n}^{t-s}$. The cascade

produces the nested index sets $\mathcal{C}_{k,0}^t \supseteq \mathcal{C}_{k,1}^t \supseteq \mathcal{C}_{k,2}^t \supseteq \mathcal{C}_{k,3}^t$, corresponding to object-associated, forward–backward-consistent, epipolar-consistent, and FoE-consistent tracks, respectively.

2.2 Track Validation and Rotation Compensation

Optical-flow tracks are first validated by forward–backward consistency [7]. For $n \in \mathcal{C}_{k,0}^t$, $\mathbf{x}_{k,n}^{t-s,b}$ denotes the point obtained by tracking $\mathbf{x}_{k,n}^{t-s}$ to I^t and back to I^{t-s}. The forward–backward error is

$$e_{k,n}^{\mathrm{FB}} = \left\| \mathbf{x}_{k,n}^{t-s} - \mathbf{x}_{k,n}^{t-s,b} \right\|_{\infty}, \qquad \mathcal{C}_{k,1}^t = \left\{ n \in \mathcal{C}_{k,0}^t \mid e_{k,n}^{\mathrm{FB}} \le \epsilon_{\mathrm{FB}} \right\}. \tag{1}$$

The surviving correspondences satisfy the calibrated epipolar constraint

$$(\widetilde{\mathbf{x}}_{k,n}^t)^{\top} \mathbf{K}^{-\top} \mathbf{E}^t \mathbf{K}^{-1} \widetilde{\mathbf{x}}_{k,n}^{t-s} = 0, \qquad \mathbf{E}^t = [\mathbf{T}^t]_{\times} \mathbf{R}^t, \tag{2}$$

where $\mathbf{R}^t$ and $\mathbf{T}^t$ denote the relative rotation and translation between frames $t-s$ and t [2]. The essential matrix is estimated from $\mathcal{C}_{k,1}^t$ using the calibrated five-point algorithm within RANSAC [8,9], and its inliers form $\mathcal{C}_{k,2}^t$. For an accepted rotation estimate, the current-frame feature position is rotation compensated as

$$\overline{\mathbf{x}}_{k,n}^t = \pi\left(\mathbf{K} (\widehat{\mathbf{R}}^t)^{\top} \mathbf{K}^{-1} \widetilde{\mathbf{x}}_{k,n}^t \right), \qquad \overline{\mathbf{v}}_{k,n}^t = \overline{\mathbf{x}}_{k,n}^t - \mathbf{x}_{k,n}^{t-s}, \tag{3}$$

where $\overline{\mathbf{x}}_{k,n}^t$ and $\overline{\mathbf{v}}_{k,n}^t$ are the rotation-compensated pixel coordinate and flow vector, respectively, and $\pi(\cdot)$ denotes dehomogenisation. If the rotation estimate is not accepted, $\widehat{\mathbf{R}}^t = \mathbf{I}$ and $\mathcal{C}_{k,2}^t = \mathcal{C}_{k,1}^t$.

2.3 Robust Object-Specific FoE Estimation

After rotation compensation, translation-induced flow associated with a rigid object should be consistent with a common FoE. Each correspondence $n \in \mathcal{C}_{k,2}^t$ defines the homogeneous flow line $\boldsymbol{\ell}_{k,n}^t = \widetilde{\mathbf{x}}_{k,n}^{t-s} \times \widetilde{\overline{\mathbf{x}}}_{k,n}^t$, which ideally passes through the object-specific FoE as illustrated in Figure 2. A RANSAC FoE hypothesis $\mathbf{f}_k^{t,(h)}$ is generated from two non-

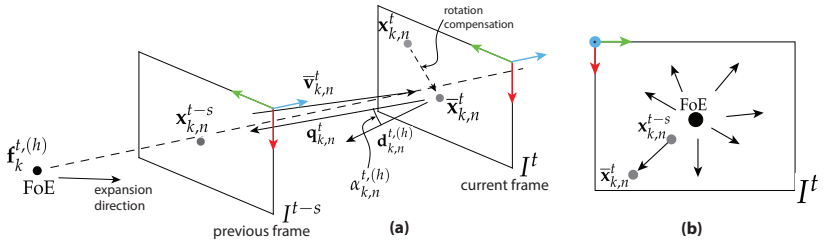


Figure 2: Rotation-compensated FoE geometry. (a) flow-line construction and angular residual $\alpha_{k,n}^{t,(h)}$. (b) collision-bound flow radiating from a common FoE.

parallel flow lines, where h indexes the hypothesis. We define the measured backward-flow direction and the corresponding ideal FoE direction as $\mathbf{q}_{k,n}^{t} = \mathbf{x}_{k,n}^{t-s} - \bar{\mathbf{x}}_{k,n}^{t}$, $\mathbf{d}_{k,n}^{t,(h)} = \mathbf{f}_{k}^{t,(h)} - \bar{\mathbf{x}}_{k,n}^{t}$. Their angular deviation is

$$\alpha_{k,n}^{t,(h)} = \arccos\left(\frac{(\mathbf{q}_{k,n}^{t})^{\top}\mathbf{d}_{k,n}^{t,(h)}}{\|\mathbf{q}_{k,n}^{t}\|_2\|\mathbf{d}_{k,n}^{t,(h)}\|_2}\right). \tag{4}$$

The three residual models; perpendicular, circular, and square are $e_{D,k,n}^{t,(h)} = \|\bar{\mathbf{v}}_{k,n}^{t}\|_2 \sin\alpha_{k,n}^{t,(h)}$, $e_{C,k,n}^{t,(h)} \approx \frac{\|\bar{\mathbf{v}}_{k,n}^{t}\|_2}{2\Delta r}\alpha_{k,n}^{t,(h)}$, and $e_{S,k,n}^{t,(h)} = \frac{\alpha_{k,n}^{t,(h)}}{\delta\alpha_{S,k,n}^{t,(h)}}$, where Δr is the circular endpoint-uncertainty radius and $\delta\alpha_{S,k,n}^{t,(h)}$ is the maximum angular perturbation induced by the square pixel grid. Their derivations are given in [10, 11]. The perpendicular residual measures pixel deviation from radial flow, whereas the circular and square models normalise directional error using circular and pixel-grid endpoint uncertainty, respectively. The latter therefore account for the greater directional uncertainty of short flow vectors. For residual model $m \in \{D, C, S\}$, the consensus set for hypothesis h is

$$\mathcal{I}_{k,m}^{t,(h)} = \left\{ n \in \mathcal{C}_{k,2}^{t} \mid e_{m,k,n}^{t,(h)} \le \epsilon_m \right\}, \quad h_m^{\star} = \arg\max_{h} \left|\mathcal{I}_{k,m}^{t,(h)}\right|. \tag{5}$$

Here, ϵ_D is specified in pixels, while the normalised circular and square residuals use $\epsilon_C = \epsilon_S = 1$. The final set is $\mathcal{C}_{k,3}^{t} = \mathcal{I}_{k,m}^{t,(h_m^{\star})}$, from which $\widehat{\mathbf{f}}_{k}^{t}$ is refined by SVD least squares.

2.4 Feature- and Object-Level TTC

For an approaching object, $\tau_k = -D_k/\dot{D}_k$, $\dot{D}_k < 0$, where τ_k is the TTC, D_k the current object distance, and $\dot{D}_k$ its rate of change. Under perspective projection, metric scale cancels and TTC is recovered from image expansion [5, 11]. For $n \in \mathcal{C}_{k,3}^t$, feature-to-FoE radial distance $\rho_{k,n}^{t-s}$ and rotation-compensated flow magnitude $g_{k,n}^t$ are defined as

$$\rho_{k,n}^{t-s} = \left\| \mathbf{x}_{k,n}^{t-s} - \widehat{\mathbf{f}}_k^t \right\|_2, \quad g_{k,n}^t = \left\| \overline{\mathbf{v}}_{k,n}^t \right\|_2, \quad \Rightarrow \quad \widehat{\tau}_{k,n}^t = \Delta t_s \frac{\rho_{k,n}^{t-s}}{g_{k,n}^t}. \tag{6}$$

The TTC-valid set $\mathcal{V}_k^t \subseteq \mathcal{C}_{k,3}^t$ and final object estimate are

$$\mathcal{V}_k^t = \left\{ n \in \mathcal{C}_{k,3}^t \;\middle|\; g_{k,n}^t > \epsilon_v,\ \rho_{k,n}^{t-s} > \epsilon_\rho,\ \frac{(\overline{\mathbf{v}}_{k,n}^t)^\top (\overline{\mathbf{x}}_{k,n}^t - \widehat{\mathbf{f}}_k^t)}{g_{k,n}^t \left\| \overline{\mathbf{x}}_{k,n}^t - \widehat{\mathbf{f}}_k^t \right\|_2} \geq \epsilon_{\text{rad}} \right\}, \tag{7}$$
$$\widehat{\tau}_k^t = \text{median}_{n \in \mathcal{V}_k^t} \widehat{\tau}_{k,n}^t.$$

The validity test rejects weak, near-FoE, and contracting motion; an object estimate is accepted only when the required minimum number and fraction of valid features are available.

3 Experimental Methodology

3.1 Experimental Setup

Simulation uses a Gazebo X3 quadrotor with a forward-facing 640×480, 30-Hz monocular camera, zero distortion, and $[f_x, f_y, c_x, c_y] = [565.6008, 565.6009, 320, 240]$ px. ROS timestamps provide $\Delta t_s = t_t - t_{t-s}$. The vehicle approaches stationary targets at 0.2, 0.5, and 1.0 $\mathrm{m\,s^{-1}}$; Gazebo motion data provide ground truth. The Gazebo-based X3 UAV simulation and data-capture software are publicly available [12].

Real validation uses an intrinsically calibrated 30-Hz DJI Tello camera and OptiTrack pose data at approximately 100 Hz. Camera and marker frames are related by extrinsic calibration, with timestamps used for temporal alignment. The target distance D_t is obtained from measured forward displacement relative to the known initial distance. OptiTrack is used only for ground truth.

3.2 Ground Truth and Evaluation Metrics

For simulated constant-velocity motion, the target distance and reference TTC at frame i are $D_i = D_0 - (d_i - d_0)$, $\tau_{\mathrm{GT},i} = \frac{D_i}{v_c}$, where d_i is the recorded displacement and v_c is estimated from the displacement–time trajectory. For an image pair $(i-s, i)$, the visual estimate is referenced to frame i and compared with $\tau_{\mathrm{GT},i}$. For the real flight, the vehicle velocity is non-uniform. The reference TTC is therefore evaluated over the same finite interval as the visual estimate,

$$\tau_{\mathrm{GT},i}^{(s)} = \Delta t_s \frac{D_i}{D_{i-s} - D_i}, \qquad \Delta t_s = t_i - t_{i-s}, \tag{8}$$

where D_i is obtained from the synchronized OptiTrack trajectory. This pair-matched reference avoids assuming constant velocity during the Tello approach. For compactness, $\tau_{\mathrm{GT},i}$ denotes the applicable ground-truth TTC for each experiment in the following evaluation metrics. For N valid estimates, let $\widehat{\tau}_i$ denote the object-level estimate for evaluation pair i and $e_i = \widehat{\tau}_i - \tau_{\mathrm{GT},i}$. TTC accuracy is evaluated by mean absolute error (MAE), root-mean-square error (RMSE), bias, median absolute error (MedAE), and mean absolute percentage error (MAPE);

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|e_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}e_i^2}, \qquad \mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}e_i, \\ \mathrm{MedAE} &= \underset{i}{\mathrm{median}}\,|e_i|, \quad \mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{e_i}{\tau_{\mathrm{GT},i}}\right|. \end{aligned} \tag{9}$$

The valid-estimate rate and stage-wise correspondence retention, $r_q = |\mathcal{C}_{k,q}^t|/|\mathcal{C}_{k,q-1}^t|$, $q \in \{1,2,3\}$ are also reported. Here, $\bar{N}$ denotes the mean number of retained correspondences per image pair.

4 Results and Discussion

4.1 Baseline TTC Accuracy and Effect of Approach Velocity

The baseline experiment considers a central approach toward a stationary human target in Gazebo. Table 1 reports the best common-window

result for each measured velocity. The preferred stride decreases with increasing speed, from $s = 32$ at 0.200 $\mathrm{m\,s^{-1}}$ to $s = 12$ at 0.490 $\mathrm{m\,s^{-1}}$ and $s = 4$ at 0.850 $\mathrm{m\,s^{-1}}$. The first two sequences achieve MAPEs of 2.33% and 2.81%, whereas the highest-speed sequence remains less accurate despite a 100% valid-estimate rate.

Table 1: Best TTC accuracy obtained for each Gazebo approach velocity.

v_c	$s^\star$	Valid [%]	MAE [s]	RMSE [s]	MAPE [%]	Bias [s]
0.200	32	100.0	0.338	0.489	2.333	0.113
0.490	12	100.0	0.171	0.274	2.811	-0.025
0.850	4	100.0	0.706	0.978	18.644	-0.036

This trend follows the image-expansion relation [13] $\eta_s \simeq \frac{v_c s}{f_{cap} D} = \frac{s}{f_{cap}\tau}$, with $\Delta t_s \simeq s/f_{cap}$. Increasing velocity or frame separation improves translational-flow observability, so high FoE consensus is reached at smaller strides as velocity increases. Larger baselines are not always beneficial. Accuracy plateaus around $s = 32$–40 at 0.200 $\mathrm{m\,s^{-1}}$, is minimized near $s = 12$ at 0.490 $\mathrm{m\,s^{-1}}$, and degrades beyond $s = 4$ at 0.850 $\mathrm{m\,s^{-1}}$. Sufficient image expansion is therefore required, while excessive inter-frame motion introduces tracking and finite-baseline effects, motivating velocity-dependent stride selection.

4.2 Contribution of the Cascaded Rejection Stages

The cascade was ablated under straight and 5° yaw-perturbed motion at approximately 0.5 $\mathrm{m\,s^{-1}}$, using the square FoE model and $s = 12$. The cumulative variants are: A_0, $\mathcal{C}_0$ with non-robust FoE estimation and mean TTC; A_1, +FB validation; A_2, +epipolar filtering and rotation compensation; A_3, +FoE RANSAC; and A_4, +median aggregation. All variants were valid for every evaluated pair (Table 2). Under straight motion, FB validation provides the main improvement, whereas rotation compensation increases MAE from 0.1286 to 0.1878 s. FoE RANSAC has negligible effect, and median aggregation mainly reduces bias. Under yaw, rotation compensation becomes critical, reducing MAE from 0.7116 to 0.3201 s; median aggregation yields the lowest final MAE of 0.2944 s. Overall, A_4 reduces yaw-case MAE by 80.8% relative to A_0.

The cascade is therefore failure-mode dependent: FB validation re-

Table 2: Cascaded ablation for straight and yaw-perturbed simulated approaches at approximately 0.5 $\mathrm{m\,s^{-1}}$ and $s = 12$.

Motion	Variant	MAE [s]	RMSE [s]	MedAE [s]	MAPE [%]	Bias [s]	$\bar{N}$
Straight	A_0 Baseline	0.1674	0.2593	0.1136	2.865	0.1149	115.2
	A_1: A_0 + FB	**0.1286**	**0.2135**	**0.0913**	**2.111**	0.0407	109.7
	A_2: A_1 + Rotation	0.1878	0.3091	0.0984	3.090	0.0365	109.1
	A_3: A_2 + FoE RANSAC	0.1874	0.3074	0.0984	3.084	0.0383	108.4
	A_4: A_3 + Median	0.1709	0.2738	0.1214	2.811	**-0.0247**	108.4
Yaw	A_0 Baseline	1.5360	2.2469	1.1298	23.876	-1.5111	112.5
	A_1: A_0 + FB	0.7116	1.6266	**0.1374**	10.324	-0.6536	92.2
	A_2: A_1 + Rotation	0.3201	**0.4932**	0.1937	5.034	**-0.0672**	90.9
	A_3: A_2 + FoE RANSAC	0.3221	0.4994	0.1937	5.049	-0.0623	89.9
	A_4: A_3 + Median	**0.2944**	0.5330	0.1506	**4.515**	-0.1412	89.9

jects unstable tracks, rotation compensation addresses rotational flow, FoE RANSAC enforces geometric consistency, and median aggregation suppresses residual TTC outliers.

4.3 Comparison of FoE Residual Models

Table 3 shows the comparison of the perpendicular, circular, and square FoE residuals on the same 0.5 $\mathrm{m\,s^{-1}}$ human-target sequence under weak ($s = 2$) and well-conditioned ($s = 12$) temporal baselines. All methods used the same current-frame interval, detections, correspondences, ground truth, and pipeline parameters.

Table 3: FoE residual-model comparison under weak ($s = 2$) and well-conditioned ($s = 12$) temporal baselines.

Stride	Residual	MAE [s]	RMSE [s]	MedAE [s]	MAPE [%]	Bias [s]	$\mathcal{C}_3/\mathcal{C}_2$ [%]
2	Perpendicular	0.8391	1.3870	0.3103	14.203	-0.7507	**95.18**
2	Circular	**0.8179**	**1.3611**	**0.2793**	**13.842**	**-0.7356**	90.85
2	Square	0.8434	1.3861	0.2956	14.293	-0.7585	91.36
12	Perpendicular	**0.1704**	0.2734	**0.1177**	**2.803**	-0.0251	**99.78**
12	Circular	0.1718	**0.2731**	0.1191	2.829	-0.0253	98.71
12	Square	0.1709	0.2738	0.1214	2.811	**-0.0247**	99.39

At $s = 2$, the residual formulation measurably affects TTC accuracy. The circular model gives the lowest MAE, RMSE, MedAE, and MAPE (0.8179 s, 1.3611 s, 0.2793 s, and 13.842%), with MAE about 2.5% below the perpendicular model and 3.0% below the square model. The perpendicular residual retains more correspondences (95.18% versus 90.85% and 91.36%), while the pixel-aware models are more selective under weak image expansion. At $s = 12$, all three models achieve

Table 4: Real-world Tello ablation at $s = 12$.

Variant	MAE [s]	RMSE [s]	MedAE [s]	MAPE [%]	Bias [s]	N
A_0 Baseline	0.7424	1.1771	0.4541	9.551	0.7310	198.5
A_1: A_0 + FB	0.7478	1.1832	0.4397	9.641	0.7432	192.9
A_2: A_1 + Rotation	0.7090	1.3867	0.2474	7.892	0.6062	189.1
A_3: A_2 + FoE RANSAC	0.7106	1.3910	0.2475	7.906	0.6085	186.9
A_4: A_3 + Median	**0.6042**	1.2587	**0.2105**	**6.566**	**0.4160**	186.9

nearly identical accuracy and retain approximately 99% of $\mathcal{C}_{k,2}^{t}$. Residual choice is therefore most consequential near the observability limit.

4.4 Real-World Validation and Ablation

The DJI Tello experiment used the square residual with $s = 12$. All 70 evaluated pairs produced valid TTC estimates. The complete pipeline achieved 0.604 s MAE, 0.211 s MedAE, and 6.566% MAPE. As shown in Figure 3(a), the largest errors occur early in the approach and decrease with increasing image expansion, consistent with Sec. 4.1. Correspon-

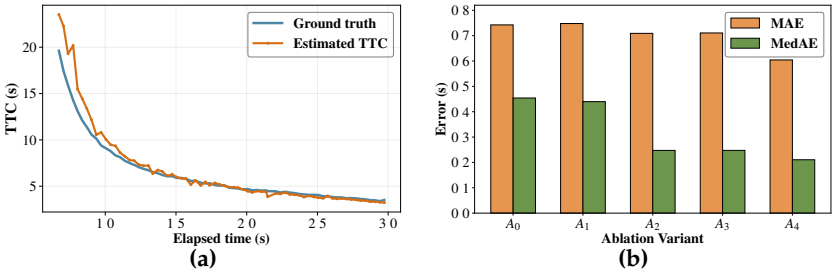


Figure 3: Real-world DJI Tello validation at $s = 12$: (a) estimated TTC and OptiTrack reference; (b) MAE and MedAE across the cumulative ablation variants.

dence retention remained high, with $\mathcal{C}_1/\mathcal{C}_0 = 97.18\%$, $\mathcal{C}_2/\mathcal{C}_1 = 98.00\%$, and $\mathcal{C}_3/\mathcal{C}_2 = 98.85\%$ ($\mathcal{C}_3/\mathcal{C}_0 = 94.18\%$). Thus, early TTC errors arise mainly from weak motion observability rather than correspondence loss.

The same 70 pairs were evaluated using the A_0–A_4 cumulative ablation defined in Sec. 4.2. The ablation in Table 4 and Figure. 3(b) shows little change from A_0 to A_1, whereas rotation compensation substantially reduces MedAE, confirming the influence of quadrotor

attitude motion. FoE RANSAC produces negligible additional change, consistent with the 98.85% $\mathcal{C}_3/\mathcal{C}_2$ retention. Median aggregation provides the strongest final improvement, reducing MAE from 0.7106 to 0.6042 s. Relative to A_0, the complete pipeline reduces MAE by 18.6% and MedAE by 53.6%. Thus, rotation compensation and robust aggregation provide the principal gains in the real sequence.

5 Conclusion

This work presented a cascaded object-specific monocular TTC framework that progressively rejects tracking-, motion-, and FoE-inconsistent correspondences before robust aggregation. The cascade reduced MAE by 80.8% under yaw-perturbed motion and transferred successfully to real DJI Tello imagery, achieving 0.604 s MAE. The experiments further show that the effectiveness of individual rejection stages depends on the underlying failure mode, while sufficient temporal baseline is required for reliable image-motion observability.

Acknowledgements

This work was partially supported by the THWS Campus for Applied Research (CAF) through the Hightech Agenda Bayern (HTA) program and by the Petroleum Technology Development Fund (PTDF), Nigeria, under the Overseas Scholarship Scheme, Ref. PTDF/ED/OSS/PHD/MI/1486/19 (19PHD058).

References

1. F. Gomes, T. Hormigo, and R. Ventura, "Vision-based real-time obstacle avoidance for drones using a time-to-collision estimation approach," in *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2020, pp. 90–95.
2. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, UK: Cambridge University Press, 2004.
3. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, vol. 2, 1981, pp. 674–679.

4. W. E. Green and P. Y. Oh, "Optic-flow-based collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 15, no. 1, pp. 96–103, 2008.
5. M. J. Iorpenda, E. Asah, and V. Willert, "Object-specific time-to-collision estimates from monocular vision for drones," in *2024 22nd International Conference on Research and Education in Mechatronics (REM)*. IEEE, 2024, pp. 225–230.
6. J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1994, pp. 593–600.
7. Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward–backward error: Automatic detection of tracking failures," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2756–2759.
8. D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.
9. M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
10. M. Buczko and V. Willert, "Monocular outlier detection for visual odometry," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 739–745.
11. A. Sabotinov, "Camera-based calculation of the time-to-collision based on the optical flow for a quadrocopter," Bachelor's thesis, Technical University of Applied Sciences Würzburg-Schweinfurt, 2025.
12. M. J. Iorpenda, "TTC Gazebo Drone," 2026. [Online]. Available: https://github.com/PendaJ/ttc-gazebo-drone
13. T. A. Camus, "Calculating time-to-contact using real-time quantized optical flow," National Institute of Standards and Technology, Tech. Rep. NISTIR 5609, 1995.

Beschreibung großer Bilddatensätze in natürlicher Sprache mit Vision-Language Models für das automatisierte Fahren

Natural-language description of large image datasets using Vision-Language Models for automated driving

Julian Truetsch[1,2], Frank Bieder[1,2] und Christoph Stiller[2]

[1] FZI Forschungszentrum Informatik, Haid-und-Neu-Str. 10, 76131 Karlsruhe
[2] Karlsruher Institut für Technologie, Kaiserstraße 12, 76131 Karlsruhe

Zusammenfassung Moderne Bildverarbeitungssysteme, insbesondere im automatisierten Fahren, werden zunehmend datengetrieben entwickelt. Das Verständnis der zugrundeliegenden, oft Millionen Bilder umfassenden Datensätze entzieht sich jedoch der manuellen Inspektion und ist über Metadaten und Labelstatistiken nur eingeschränkt zugänglich. Wir stellen einen praxisorientierten Ansatz vor, mit dem sich Bilddatensätze mit Hilfe von open-weight VLMs automatisiert in natürlicher Sprache zusammenfassen lassen. Neben der konzeptionellen Beschreibung des zweistufigen, auch für sehr große Datensätze geeigneten Proposer-Filter-Ansatzes adressiert der Beitrag auch praxisrelevante Aspekte wie Prompt-Design, GPU-Speicherbedarf, Quantisierung und strukturierte Modellausgabe mittels *guided decoding*.

Schlüsselwörter Fahrzeuge, Vision-Language Models

Abstract Modern computer vision systems, particularly in automated driving, are increasingly developed in a data-driven manner. However, understanding the underlying datasets, which often comprise millions of images, is beyond the reach of manual inspection and only partially accessible through metadata and label statistics. We present a practice-oriented approach for automatically summarizing image datasets in natural language using open-weight VLMs. In addition to a conceptual description of

the two-stage proposer-filter approach, which is also suitable for very large datasets, this paper addresses practical aspects such as prompt design, GPU memory requirements, quantization, and structured model output via *guided decoding*.

Keywords Vehicles, Vision-Language Models

1 Motivation

Mit der zunehmenden Datenorientierung in der industriellen Bildverarbeitung wird die Zusammensetzung der Trainings- und Validierungsdatensätze zu einem zentralen Qualitätsfaktor. Insbesondere Datensätze für das automatisierte Fahren sind mittlerweile zu groß, um diese noch effektiv manuell sichten zu können. In der Praxis erfolgt die Analyse großer Bilddatensätze daher überwiegend über Metadaten (Zeitstempel, GNSS-Position, Sensorkonfiguration) und vordefinierte Labels. Beide Quellen sind durch ihren festgelegten Wertebereich beschränkt und können semantische Eigenschaften, wie etwa typische Umgebungsbedingungen, häufige Objekteigenschaften oder unbeabsichtigte Biases, nur teilweise erfassen.

Aktuelle Vision-Language Models (VLMs) bieten einen alternativen Zugang: Sie können Bildinhalte in natürlicher Sprache beschreiben und sind nicht an einen vordefinierten Labelraum gebunden. Während vorhergehende Arbeiten zur *set difference captioning* [1,2] diese Fähigkeit genutzt haben, um *Unterschiede* zwischen zwei Bildmengen in natürlicher Sprache zu beschreiben, betrachten wir hier den verwandten, aber eigenständigen Fall der Beschreibung eines *einzelnen* Datensatzes.

2 Vision-Language Models und Self-Hosting

VLMs verbinden die Sprachfähigkeiten großer Sprachmodelle (LLMs) mit der visuellen Verarbeitung von Bildencodern und können textuelle Anweisungen (*Prompts*) gemeinsam mit einem oder mehreren Bildern als Eingabe verarbeiten [3]. Neben kommerziellen Modellen, die nur über kostenpflichtige APIs zugänglich sind, hat sich in den letzten Jahren eine leistungsfähige Familie an *open-weight* Modellen eta-

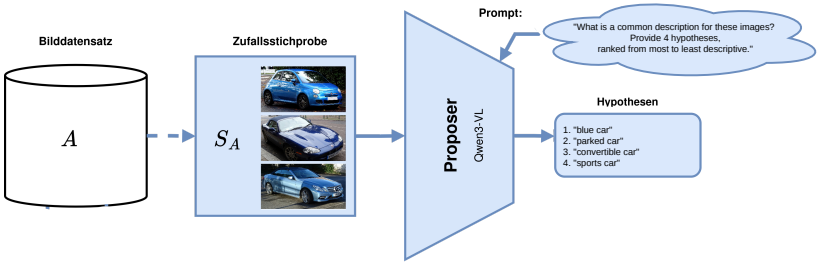


Abbildung 1: Einstufiger Ansatz zur Beschreibung eines Bilddatensatzes A in natürlicher Sprache. Der *Proposer* generiert gerankte Beschreibungshypothesen aus zufälligen Teilmengen $S_A \subset A$.

bliert, deren Gewichte frei heruntergeladen und lokal ausgeführt werden können. In unserer Implementierung verwenden wir `Qwen3-VL` [4] als VLM und LLM und `gpt-oss` [5] als reines Sprachmodell für Evaluierungszwecke. Die lokale Ausführung ist insbesondere dann wichtig, wenn sensitive Daten, wie sie in vielen industriellen Aufnahmen vorliegen, aus regulatorischen oder wettbewerblichen Gründen das Unternehmen nicht verlassen dürfen.

Für die effiziente Ausführung dieser Modelle wurde eine Reihe verschiedener Frameworks entwickelt; wir verwenden in diesem Beitrag das weit verbreitete und auch für kommerzielle Anwendungen geeignete Framework `vLLM` [6]. Es unterstützt eine Vielzahl gängiger Modellarchitekturen sowie Optimierungen wie *continuous batching* und *PagedAttention*, die für einen hohen Durchsatz bei der Verarbeitung großer Bildmengen erforderlich sind. Die Modelle lassen sich mit wenigen Zeilen Code über eine OpenAI-kompatible HTTP-Schnittstelle oder direkt über die Python-API ansprechen, was die Integration in bestehende Datenpipelines deutlich vereinfacht.

3 Beschreibung von Bilddatensätzen

Die direkte Verarbeitung aller Bilder eines Datensatzes durch ein VLM ist durch das begrenzte Kontextfenster aktueller VLMs nicht möglich. Stattdessen wird (ggf. wiederholt) eine zufällige Teilmenge $S_A \subset A$ einer festen Größe gezogen und zusammen mit einem Prompt,

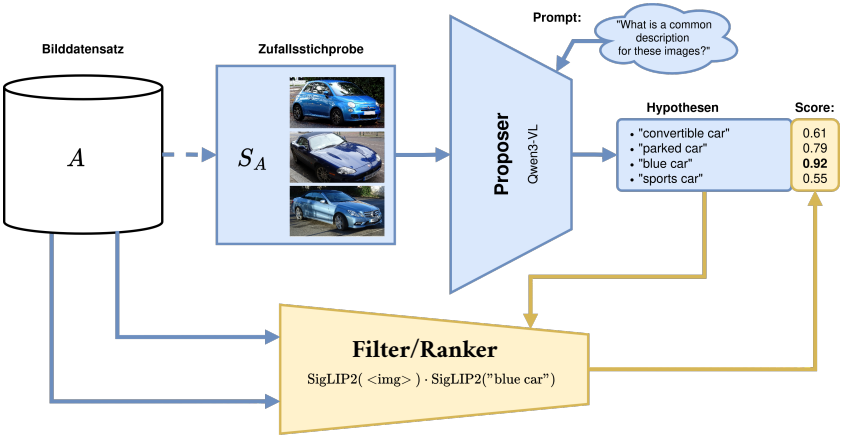


Abbildung 2: Zweistufiger Ansatz zur Beschreibung eines Bilddatensatzes A in natürlicher Sprache. Der *Proposer* generiert Beschreibungshypothesen aus zufälligen Teilmengen $S_A \subset A$; der *Filter/Ranker* verifiziert für jede Hypothese, auf welchen Anteil des vollständigen Datensatzes A sie zutrifft.

der zur Generierung mehrerer Beschreibungshypothesen auffordert, an das VLM übergeben. Mehrere Wiederholungen liefern einen Pool an Kandidaten-Beschreibungen. Alternativ können Bildbeschreibungen (Captions) zunächst einzeln erzeugt und anschließend von einem reinen LLM zu Hypothesen aggregiert werden.

Bei einem solchen stichprobenbasierten Ansatz stellt sich die Frage, ob die auf einer einzelnen Bilder-Stichprobe erhaltenen Beschreibungen tatsächlich auf den Datensatz in seiner Gesamtheit zutreffen, oder nur das Resultat von Sampling-Bias sind. Je diverser der Datensatz, desto schwieriger ist es, den Datensatz durch eine einzelne Stichprobe adäquat abzubilden. Das Risiko, durch Sampling-Bias verzerrte Stichproben zu erhalten steigt zudem unter Rauschen [1, 2], also wenn einzelne Bilder *nicht* zu einer übergreifenden Datensatzbeschreibung passen, unabhängig davon, ob es sich dabei um natürliche Ausreißer oder Fehler bei der Zusammenstellung des Datensatzes handelt.

Wir untersuchen daher neben einem einstufigen, auf einer einzelnen Stichprobe basierenden Ansatz (Abb. 1), auch zweistufige Architekturen (Abb. 2), bestehend aus einem *Proposer*, der, wie zuvor beschrie-

ben, ausgehend von zufälligen Teilmengen Beschreibungshypothesen generiert, sowie einem nachgelagerten *Ranker* oder *Filter*, der diese Hypothesen über den vollständigen Datensatz verifiziert. Der einstufige Ansatz gibt hingegen direkt die auf der Stichprobe vom Proposer-VLM generierten Hypothesen als Datensatzbeschreibung aus.

3.1 Ranker

Der zweistufige Ansatz mit einem Ranker ist die direkte Adaption der VisDiff-*set difference captioning*-Methodik [1,2], wird hier jedoch auf den Ein-Datensatz-Fall übertragen und entsprechend angepasst.

Der Ranker berechnet für jede Hypothese eine Metrik, die beschreibt, wie gut diese Hypothese den Datensatz als Gesamtes beschreibt. Entscheidend dabei ist, dass diese Metrik alle Bilder des Datensatzes berücksichtigt, und nicht nur die Bilder der Proposer-Stichprobe. Die Hypothesen werden dann entsprechend dieser Metrik sortiert, und die top-N-Hypothesen werden als Datensatzbeschreibung ausgegeben. Im einfachsten Fall wird jedes Bild einzeln von einem VLM mit der Hypothese verglichen, ein Score berechnet wie gut die Hypothese zu dem Bild passt, und dieser Score über den gesamten Datensatz gemittelt.

Wir betrachten zunächst zwei Varianten, die beide direkte Adaptionen des VisDiff-Rankers [1] für Beschreibungen eines einzelnen Datensatzes sind: (i) ein generativer Ranker, der das VLM fragt, ob die jeweilige Hypothese zutrifft; (ii) ein eingebettungsbasierter Ranker mit `SigLIP 2` [7], der Hypothese und Bilder in einen gemeinsamen Vektorraum projiziert und die Kosinusähnlichkeit als Score interpretiert.

3.2 Experimente

Wir evaluieren sowohl den einstufigen als auch die zweistufigen Ansätze auf dem AD-Diff-Bench-Benchmark [2]. AD-Diff-Bench ist ursprünglich für die Aufgabe *set difference captioning* [1] konzipiert, bei der Unterschiede zwischen Bilder-Set-Paaren gefunden werden sollen. Jedes Set-Paar des Benchmark-Datensatzes verfügt bereits über eine Beschreibung des ersten Sets des Paares, wir verwenden für unsere Evaluation entsprechend jeweils nur das erste Set eines Set-Paares.

Der Benchmark-Datensatz besteht aus drei Splits (Abb. 3): `web-scraped`, `annotation-filtered` und `CLIP-filtered`. Der Split

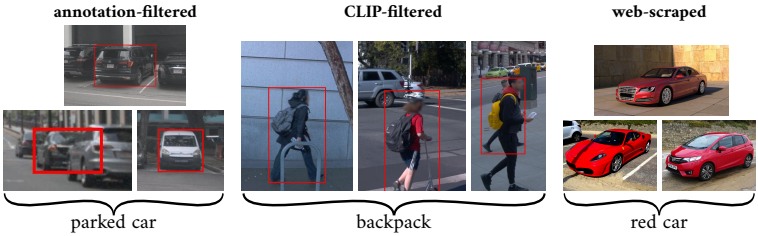


Abbildung 3: Beispielbilder mit Ground-Truth-Beschreibung. jeweils eines zu beschreibenden Bilder-Sets für jeden der drei Splits (web-scraped, annotation-filtered, CLIP-filtered) des Benchmark-Datensatzes AD-Diff-Bench [2].

Tabelle 1: Einstufige und zweistufige Verfahren ohne Domänen-Referenz, acc@1 (Mittel über drei Seeds, Streuung $\leq .039$). *Einstufig* (B2): VLM ohne Ranker (1×100 Bilder); *Zweistufig*: VLM oder LLM auf Captions als Proposer (3×20 Bilder, Qwen3-VL-30B-A3B), dann Ranker (mittlere SigLIP2-Kosinusähnlichkeit zur Zielmenge (SigLIP2-Ranker, B1 und B3); oder gemittelte VLM-Bewertung per Bild (VLM-Ranker, B4, Qwen3.5-35B-A3B, nur annotation-filtered)).

		annotation-filtered (60)	web-scraped (180)
B2	einstufig	**.625**	**.604**
B1	zweistufig (VLM) + SigLIP2-Ranker	.447	**.599**
B3	zweistufig (LLM) + SigLIP2-Ranker	.358	**.613**
B4	zweistufig (VLM) + VLM-Ranker	.431	–

`annotation-filtered` besteht aus Bildern aus Datensätzen für das automatisierte Fahren (KITTI [8], NuImages [9], Waymo Open Perception [10]), gefiltert in unterschiedliche Sets durch Datensatz-Label und Metadaten. `CLIP-filtered` basiert auf denselben Datensätzen, verwendet aber CLIP [11], um Bilder anhand ihrer Kosinusähnlichkeit zu vordefinierten Schlagwörtern im CLIP-Embeddingraum in unterschiedliche Sets zu sortieren, die Ground-Truth-Beschreibung ist dann das jeweilige Schlagwort. `web-scraped` basiert auf mittels Bing-Bildersuche aus dem Internet gescrapten Bildern, der jeweilige Suchbegriff definiert in diesem Fall das Ground-Truth-Label.

Unsere Evaluation basiert nur auf den Splits `annotation-filtered` (`ann60`) und `web-scraped`, da `CLIP-filtered` zu Gunsten des SigLIP2-basierten Rankers und Filters verzerrt ist. Sowohl bei CLIP als auch

bei SigLIP2 handelt es sich um kontrastiv auf Web-Alt-Texten trainierte Dual-Encoder. Diese Ähnlichkeit in Architektur und Trainingsdaten lässt geteilte Fehlermodi erwarten, weshalb eine Überschätzung der Genauigkeit eines SigLIP2-basierten Ansatzes auf einem mit CLIP konstruierten Benchmark zu befürchten ist.

Das Evaluationsprotokoll ist an das von AD-Diff-Bench angelehnt; das open-weights LLM `gpt-oss` [5] vergleicht die ausgegebene Datensatzbeschreibung mit der Ground-Truth-Beschreibung auf semantische Äquivalenz und gibt eine Bewertung von `0` (keine Übereinstimmung), `1` (teilweise Übereinstimmung) oder `2` (semantisch äquivalent) aus. Strukturierte Modellausgabe mittels *guided decoding* garantiert, dass das LLM niemals mit einer anderen Ausgabe als `0`, `1` oder `2` antworten kann. Den Prompt für gpt-oss, der ursprünglich auf einem manuell gelabelten Hypothesen-Datensatz getunt wurde, haben wir mit geringfügigen Anpassungen übernommen. Ein detaillierter, strukturierter Prompt, sowohl mit praktischen Beispielen als auch einem differenzierten Regelwerk hat sich dabei als am effektivsten herausgestellt. Proposer-VLM/LLM verwenden einen deutlich kürzeren Prompt, der neben klaren Anweisungen ebenfalls Beispiele enthält.

Sowohl für den einstufigen als auch den zweistufigen Ansatz verwenden wir in allen Experimenten `Qwen3-VL` [4] als Proposer-VLM/LLM. Im zweistufigen Ansatz mit VLM-Proposer (LLM-Proposer) werden für drei Stichproben von jeweils 20 Bildern (Captions) vom Proposer-VLM (Proposer-LLM) jeweils 10 Hypothesen generiert. Der einstufige Ansatz verwendet nur eine Stichprobe von 100 Bildern und generiert fünf Datensatzbeschreibungen. Der SigLIP2-Ranker des zweistufigen Ansatzes lädt die Modellvariante `SigLIP 2 Giant`. Der alternative VLM-Ranker basiert auf `Qwen3.5-35B-A3B` und verwendet wie `gpt-oss` guided decoding für ein garantiert eingehaltenes, automatisch parsebares Ausgabeformat.

Der zweistufige Ansatz liefert auf `annotation-filtered` deutlich schlechtere Ergebnisse als der einstufige Ansatz (Tab. 1). Auf `web-scraped` erzielen dagegen alle getesteten Ansätze ähnliche Ergebnisse. Es ist anzunehmen, dass dies an dem Domain-Shift zwischen web-scraped (Bilder aus dem Internet) und annotation-filtered (Straßenverkehr) liegt: SigLIP2 ist auf aus dem Internet gescrapten Bildern und deren Alt-Texts trainiert, und die Embeddings sind entsprechend den a-priori Wahrscheinlichkeiten von Bildern und Texten im Inter-

net angepasst, nicht spezifisch auf die Domäne Straßenverkehr. Der LLM-Proposer ist auf `annotation-filtered` deutlich schlechter als der VLM-Proposer, und vergleichbar auf `web-scraped`. Im Folgenden untersuchen wir daher nur Varianten mit VLM-Proposer.

Der Einfluss von guided decoding auf den VLM-Ranker ist vernachlässigbar. Von 617.697 Ranking-Schritten *ohne* guided decoding trat nur zweimal eine Verletzung des im Prompt festgelegten Ausgabeformats auf. Der VLM-Ranker ist um mehr als einen Faktor drei langsamer als der SigLIP2-Ranker. Dies liegt darin begründet, dass generative VLMs autoregressiv arbeiten, für jedes Bild also mehrere Forward-Passes durch das Modell erfolgen müssen, anstatt nur ein Forward-Pass bei einem Embedding-Model wie SigLIP2. Da die Genauigkeit zudem leicht unter der des SigLIP2-Rankers liegt, wird im Folgenden in allen Experimenten der SigLIP2-Ranker (oder Filter) verwendet.

4 Beschreibung von Bilddatensätzen unter Domain-Shift

Wenn die schlechteren Ergebnisse des SigLIP2-Rankers auf dem `annotation-filtered` Split im Vergleich zum `web-scraped` Split im Domain-Shift von Bildern aus Datensätzen für das automatisierte Fahren zu den SigLIP2-Trainingsdaten begründet sind, bietet sich eine Standardisierung der SigLIP2-Kosinusähnlichkeiten auf einen innerhalb der Domäne gleichen Mittelwert und Varianz an: $z_H = (\bar{s}_{\text{Ziel},H} - \mu_{\text{Ref},H})/\sigma_{\text{Ref},H}$, mit Mittelwert $\mu_{\text{Ref},H}$ und Standardabweichung $\sigma_{\text{Ref},H}$ der Kosinusähnlichkeit der Hypothese H auf einer Domänen-Referenz, also einem ausreichend großen Datensatz für die Domäne repräsentativer Bilder. Für AD-Diff-Bench nutzen wir, entsprechend den zwei verschiedenen Domänen in AD-Diff-Bench (Internet-Bilder und Straßenverkehr) zwei Domänen-Referenzen, jeweils aus den Bildern des entsprechenden Datensatz-Splits (`annotation-filtered` bzw. `web-scraped`). Bilder aus der Zielmenge werden aus den Domänen-Referenzen entfernt.

Dass der SigLIP2-Ranker das Ranking von Proposer-Hypothesen in der set difference captioning Aufgabe klar verbessert [2], aber nicht für Beschreibungen einzelner Sets, motiviert eine weitere Familie an Verfahren, die die Domänen-Referenz als Referenzset im Sinne des set difference captioning nutzt. Eine Hypothese wird umso höher

gerankt, je mehr sie sich vom Domänen-Hintergrund unterscheidet. Die VisDiff- [1] und AD-Diff- [2] Ranker verwenden als Vergleichsmetrik zwischen den beiden Sets den AUROC zwischen den Kosinusähnlichkeitsverteilungen der jeweiligen Hypothese H auf beiden Sets. Wir untersuchen zudem zwei weitere Metriken für den Abstand zwischen den beiden Sets: die KL-Divergenz und die p-Werte eines Permutationstests zwischen den Kosinusähnlichkeitsverteilungen.

4.1 Filter

Ein alternativer Ansatz zum Ranking aller Proposer-generierten Hypothesen ist ein Filter-Schritt als zweite Stufe nach dem Proposer. Die Reihenfolge der vom Proposer ausgegebenen Hypothesen wird in diesem Fall nicht geändert, stattdessen werden nur diejenigen Hypothesen aussortiert, die nicht auf den Datensatz in seiner Gesamtheit zutreffen, die also Resultat von Sampling-Bias sind. Im Allgemeinen ist es schwierig, einen konkreten Schwellwert für die mittlere Kosinusähnlichkeit festzulegen, was die Anwendung dieses Verfahrens ohne Referenzdatensatz nicht erlaubt. Ist aber eine Domänen-Referenz verfügbar, so lassen sich für den Filter dieselben Metriken wie für den Ranker anwenden (AUROC, KL-Divergenz, Permutationstest).

4.2 Experimente

Tabelle 2 zeigt, dass sowohl Ranking über standardisierte mittlere Kosinusähnlichkeit, als auch Ranking über AUROC, KL-Divergenz, oder Permutations-p zwischen Zielmenge und Domänen-Referenz die Genauigkeit des Rankers auf dem Datensatz mit Domain-Shift (`annotation-filtered`) im Vergleich zum Ranking ohne Domänen-Referenz (`zweistufig (VLM)`) wesentlich verbessert, und dort nahezu die Genauigkeit des einstufigen Ansatzes erreicht. Unter 20% Domänenrauschen, wenn also 20% der Zielmenge durch zufällig aus derselben Domäne gesampelte Bilder ausgetauscht werden, verstärkt sich dieser Trend noch weiter. Einstufiger Ansatz und Domänen-Referenz-Ansätze sind dann als gleichwertig zu betrachten. Ein eindeutiger Vorteil des Rankers mit Domänen-Referenz gegenüber dem einstufigen Ansatz ergibt sich jedoch auch in diesem Fall nicht.

Tabelle 2: Ranking und Filtern der Kandidaten, ohne und mit Domänenrauschen (20% jeder Zielmenge durch Bilder aus der Domäne ersetzt). Filter behalten obere 50%, sofern nicht anders angegeben; *Permutations-p*: $p \leq 0.05$. Mittel über drei Seeds; ΔB1 und ΔB2 sind gepaarte Differenzen gegenüber B1 bzw. B2. **Fett**: Bestes Verfahren des Splits; unterstrichen: bestes im Abschnitt. B3 wurde nicht mit Rauschen ausgewertet.

		Methode	annotation-filtered (60) acc@1	ΔB1	ΔB2	web-scraped (180) acc@1	ΔB1	ΔB2
ohne Rauschen	Basel.	B2 einstufig	.625	+.178	+.000	**.604**	+.004	+.000
		B1 zweistufig (VLM)	.447	+.000	−.178	**.599**	+.000	−.004
		B3 zweistufig (LLM)	.358	−.089	−.267	**.613**	+.012	+.008
	Ranker mit Dom.-Ref.	Standardisierung	.547	+.100	−.078	.568	−.031	−.035
		AUROC	.564	+.117	−.061	.581	−.019	−.022
		KL-Divergenz	.597	+.150	−.028	.582	−.017	−.021
		Permutations-p	.539	+.092	−.086	.579	−.021	−.024
	Filter mit Dom.-Ref.	Standardisierung	**.653**	+.206	+.028	.576	−.023	−.027
		AUROC	**.667**	+.219	+.042	.575	−.024	−.028
		KL-Divergenz (> 25%)	**.675**	+.228	+.050	.581	−.019	−.022
		Permutations-p	**.647**	+.200	+.022	.577	−.022	−.026
mit 20% Rauschen	Basel.	B2 einstufig	.603	+.222	+.000	**.569**	−.003	+.000
		B1 zweistufig (VLM)	.381	+.000	−.222	**.572**	+.000	+.003
	Ranker mit Dom.-Ref.	Standardisierung	.558	+.178	−.044	**.568**	−.005	−.002
		AUROC	**.592**	+.211	−.011	.531	−.041	−.038
		KL-Divergenz	**.600**	+.219	−.003	**.568**	−.005	−.002
		Permutations-p	.550	+.169	−.053	**.571**	−.001	+.002
	Filter mit Dom.-Ref.	Standardisierung	**.642**	+.261	+.039	.568	−.005	−.002
		AUROC	**.636**	+.256	+.033	.563	−.009	−.007
		KL-Divergenz (> 25%)	**.642**	+.261	+.039	**.581**	+.008	+.011
		Permutations-p	.622	+.242	+.019	.558	−.014	−.011

Der Filter-Ansatz mit Domänen-Referenz hingegen erzielt sowohl mit als auch ohne Domänen-Rauschen die besten Ergebnisse auf `annotation-filtered`, mit einer höheren Genauigkeit als der einstufige Ansatz für alle getesteten Metriken.

Alle Domänen-Referenz-Ansätze, sowohl Ranker- als auch Filter-basiert, sind auf `web-scraped`, dem Datensatz ohne Domain-Shift, erwartungsgemäß nahezu neutral. Im Vergleich der verschiedenen Metriken zueinander erzielt KL-Divergenz die besten Ergebnisse, und Stan-

dardisierung und Permutations-p die schwächsten, sowohl für Ranker als auch Filter, mit ähnlichen Trends mit und ohne Domänenrauschen.

5 Hardwareanforderungen

Die Modelle wurden parallel mit vLLM-Docker-Containern auf fünf Nvidia RTX 6000 Ada GPUs mit jeweils 48 GiB GPU-Speicher verteilt. Das Proposer-VLM/LLM Qwen3-VL-30B-A3B (30 Mrd. Parameter) belegt unquantisiert in bf16 (ca. 2 Byte per Parameter) 59 GiB GPU-Speicher allein für die Modellgewichte, und wurde daher tensorparallel auf zwei GPUs verteilt deployed. Das Evaluations-LLM gpt-oss-120b (120 Mrd. Parameter) ist MXFP4-quantisiert und benötigt etwas mehr als ein halbes Byte pro Parameter, insgesamt 69 GiB an GPU-Speicher für Gewichte, und wird daher ebenfalls verteilt auf zwei GPUs deployed. SigLIP2-giant misst insgesamt ca. 1,8 Mrd. Parameter, 1,1 Mrd. Parameter für den Vision-Encoder und 0,7 Mrd. für den Text Encoder. Damit benötigt es in bf16 nur 3,5 GiB für die Modellgewichte, und passt auf eine GPU. Qwen3.5-35B-A3B (35 Mrd. Parameter) wurde in FP8-Quantisierung deployed, damit benötigt es bei ungefähr gleicher Parameterzahl nur etwas mehr als die Hälfte des GPU-Speichers von Qwen3-VL (34 GiB), und passt damit auf eine GPU.

Zusätzlicher GPU-Speicher wird bei allen autoregressiven Modellen für den KV-Cache benötigt, abhängig von der Anzahl an Bildern und der Länge des Textes der vom Modell gleichzeitig verarbeitet wird. Bei moderaten Kontextlängen kamen alle unsere Modelle mit wenigen GiB KV-Cache aus, auch gpt-oss als einziges Modell mit durch aktiviertes Reasoning erhöhtem KV-Cache-Bedarf. Um wiederholte Benchmark-Sweeps zu beschleunigen, wurden alle Modellausgaben (Texte und Embeddings) mittels Valkey [12] gecached.

6 Fazit

In diesem Beitrag konnten wir demonstrieren, wie sich bereits mit einem einfachen, einstufigen VLM-Ansatz große Bilddatensätze in natürlicher Sprache zusammenfassen lassen. Ein zweistufiger Proposer-Ranker-Ansatz führt, anders als bei Beschreibungen von Unterschieden zwischen Datensätzen [1, 2], für diese Anwendung im

Allgemeinen nicht zu besseren Ergebnissen. Für den in der Praxis häufigen Fall von Daten mit hohem Domain-Shift zu den Trainingsdaten der verwendeten VLMs, wie etwa bei Anwendungen für das automatisierte Fahren, führt ein zweistufiger Proposer-Filter-Ansatz dagegen zu robusteren Beschreibungen als ein einstufiger Ansatz.

Literatur

1. L. Dunlap *et al.*, "Describing differences in image sets with natural language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.
2. J. Truetsch, F. Hauser, C. Stiller, and F. Bieder, "Understanding autonomous driving datasets by describing differences between image subsets in natural language," 2026, arXiv:2609.03677.
3. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
4. S. Bai *et al.*, "Qwen3-VL technical report," 2025, arXiv:2511.21631.
5. OpenAI, "Introducing gpt-oss," OpenAI Blog. https://openai.com/index/introducing-gpt-oss/, 2025.
6. W. Kwon *et al.*, "Efficient memory management for large language model serving with PagedAttention," in *Proc. 29th Symp. Oper. Syst. Princ.*, 2023.
7. M. Tschannen *et al.*, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025, arXiv:2502.14786.
8. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, 2013.
9. H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.
10. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.
11. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
12. The Linux Foundation, "Valkey," https://valkey.io, 2026.

Optical quality and computer vision performance: Emerging new metrics

Alexander Braun

Hochschule Düsseldorf, Fachbereich Elektro- und Informationstechnik, Münsterstr. 156, 40476 Düsseldorf

Abstract An important characteristic of the optical quality of camera systems is the modulation transfer function (MTF). In industrial practice it is always condensed into a single scalar such as MTF50. That scalar was devised for human viewers and automated quality control, and the evidence of the last several years shows it to be a limited predictor of the performance of computer vision based on neural networks. While the general rule "More blur, less performance" holds for computer vision (CV) models over a large range, it fails most clearly in the regime that matters for series production: the narrow band of variation caused by manufacturing tolerances, which is where camera modules are accepted or scrapped in the millions. The alternatives are not new — the groundwork was laid in medical imaging and in photographic science decades ago — but only over the last several years has the pace picked up, and this contribution reports that development: a move from scalar MTF values towards *integral* metrics evaluating the whole frequency response and, in part, the noise spectrum. We consider four candidates under discussion, namely the Strehl ratio, the optical informative gain, the Shannon information capacity and the ideal-observer signal-to-noise ratio, show what they share, and how well they correlate with CV algorithms.

Keywords Image quality, modulation transfer function, information capacity, Strehl ratio, ideal observer, computer vision, end-of-line testing, automotive cameras

1 Introduction

Every front-looking automotive camera is a safety-relevant component built in high volume, and its output is consumed not by a human but by a learned algorithm. Throughout this contribution, computer vision (CV) means perception by a trained neural network, whether a convolutional architecture or a vision transformer. The optical quality of such a camera therefore has to be specified against a criterion that says something about how well the network will subsequently perform. The community has spent the last decade building the measurement infrastructure for that task, and IEEE 2020, formerly P2020, is now approved and published as a standard for automotive image quality [1].

One relevant metric of that infrastructure is the MTF, respectively its measured counterpart, the SFR.[1] In industrial practice the full curve is almost never used. What is used is a scalar extracted from it, most commonly MTF50 (sometimes MTF20), or the MTF at a fixed fraction of the Nyquist frequency ('at half Nyquist '). A production line needs a Boolean gate, and a Boolean gate needs a number and a limit. At end-of-line (EOL) testing, a value above the limit means the module ships with the implicit promise that it will not cause harm in the field; a value below the limit means it is reworked or, more commonly, destroyed, discarding the accumulated added value. This is where the money is [2].

The difficulty is that MTF50 was developed for a different observer. It was introduced as a correlate of perceived sharpness for a human viewer. A neural network responds to a different combination of image properties, and weighs frequency regimes differently than the human contrast sensitivity function (CSF) does. Two lenses with identical MTF50 can therefore yield different CV performance, and several studies have documented this [3–6]. The correlation between optical scalar and network performance is strong when the optical degradation is strong, and it becomes unreliable in the regime that matters for mass production [4].

An answer has been forming for a long time, and it has recently gathered pace. Its roots go back two decades in photographic science,

[1] MTF and SFR are used synonymously throughout. SFR is the more careful term whenever the measurement derives from a non-harmonic target such as a slanted edge.

and further still if the medical-imaging literature is counted, but only over the last several years has it been brought to bear on automotive cameras and on learned perception. Both the IS&T Electronic Imaging conferences and the closely related IEEE 2020 working group have proposed such metrics, and they share one structural feature, in that they are *integral* metrics. Rather than reading a single point off the MTF curve, they integrate the curve, or a noise-weighted version of it, over spatial frequency, and thereby retain information that any single-point read-out discards. Four are currently under discussion: the Strehl ratio, the optical informative gain (OIG), the Shannon information capacity, and the ideal-observer signal-to-noise ratio (SNR_i).

This contribution presents no original results, but reports this trend.

2 The Scalar at End of Line

2.1 Why Industry Requires a Scalar

The requirement for a scalar comes not from optics but from the automotive quality process. Under IATF 16949 [7] a characteristic relevant to a safety function must be specified as an inspection characteristic: a measurand with a nominal value, a tolerance interval and a documented measurement capability, applied to each part by comparing one measured value against one limit. Three requirements follow: the metric must reduce to a single number per device under test; a limit must exist, which requires monotonicity in the quantity of interest across the whole tolerance window; and it must be measurable within the tact time of an EOL station, from a chart that fits the available working distance. MTF50 satisfies the first and third, and is assumed to satisfy the second. It is that assumption which the rest of this section examines.

2.2 The Evidence That the Scalar Fails

The relationship between MTF50 and CV performance has to be stated at two different resolutions. Across a wide range of optical quality the relationship holds: Müller et al. [8] sampled 100 real lenses spanning a large range of achievable quality, defined lens quality as the field-averaged MTF50, applied the resulting point spread functions (PSF) to ImageNet and MSCOCO, and found lens quality and model accuracy

strongly correlated — a Pearson coefficient of 0.86 and a Kendall rank coefficient of 0.81 for classification, and 0.84 and 0.65 for detection, the latter at p-values close to zero.

The relationship fails at the resolution at which EOL testing operates. At a given optical quality the spread of CV performance depends on the concrete *form* of the PSF: the aberration type — astigmatism, defocus, coma — matters independently of quality. Two lenses that an EOL station would score identically can therefore differ in the aberration composition the network actually responds to, and perform differently in the field. The scalar MTF50 value does not resolve these differences.

The consequence is visible in individual cases. Müller and Braun [4] report two degradation algorithms yielding MTF values of 22 % and 40 % at 0.2 cycles/pixel — a difference that in an automotive quality process would separate a fail from a pass — with person-detection confidences of 91.3 % and 90.2 %, the lower MTF giving the marginally higher confidence.

An earlier study [3] corroborates this finding. The average precision of the CV algorithm proved robust against the level of approximation inside the simulated lens model, while the MTF depended distinctly on that level. Wolf et al. [5] reach the same conclusion from the windshield side: from the refractive power or the MTF at half Nyquist, the performance of a segmentation network cannot be inferred unambiguously. Geever et al. [6] quantify the spread on a larger sample, obtaining a Spearman rank correlation of 0.65 between MTF50 and detection accuracy, with a wide band of accuracies at any given MTF50 and a wide band of MTF50 values at any given accuracy.

A separate observation shows that the problem is not only one of the spread of the measured values. Koren [9] measured MTF50 on one and the same 24 MP camera under five ISP filter settings, from a mild low-pass filter to strong sharpening. MTF50 varies by a factor of 3.4, from 0.160 to 0.541 cycles/pixel, without any change to the optics. Since sharpening amplifies noise along with the signal, and the strongest MTF50 in that set comes from the strongest sharpening, an MTF50 limit can be met by ISP processing rather than by optics.

In summary, the evidence that MTF50 does not predict CV performance for driver assistance cameras is accumulating. The aberrations that production tolerances actually produce are not captured by MTF50 but do influence CV performance; conversely, MTF50 values differing

by a factor of three have been shown not to influence it at all.

3 Emerging New Metrics

The four metrics were derived independently and are written in four notations, which obscures a common structure: each integrates the transfer function of the system over spatial frequency rather than sampling it at one point. Sec. 3.1 gives the definitions, and Sec. 3.2 gives the properties that follow from the choices and decide where in the supply chain each metric can be used.

3.1 Definitions

Tab. 1 collects the four definitions in a common notation, together with the source each is taken from. Written this way the shared structure is visible directly: every one of them is an integral over spatial frequency of a function of the MTF, and they differ only in which function.

Table 1: Definitions of the four integral metrics in a common notation. $\mathrm{MTF}_{\mathrm{dl}}$ is the diffraction-limited reference at the same f-number and wavelength, NPS the noise power spectrum, μ the mean linear signal and G the Fourier transform of the difference between object and background. The integration limits are those of the measured case, i.e. up to the Nyquist frequency of 0.5 cycles/pixel.

Metric	Definition	Source
Strehl ratio	$\mathrm{SR} = \dfrac{\int \mathrm{MTF}\,\mathrm{d}k}{\int \mathrm{MTF}_{\mathrm{dl}}\,\mathrm{d}k}$	[5,10]
OIG	$\mathrm{OIG} = \dfrac{\int \mathrm{MTF}^2\,\mathrm{d}k}{\int \mathrm{MTF}_{\mathrm{dl}}^2\,\mathrm{d}k}$	[5]
Capacity	$C = \displaystyle\int_0^{1/2} \log_2\left(1 + \mu^2\,\frac{\mathrm{MTF}^2(f)}{\mathrm{NPS}(f)}\right)\mathrm{d}f$	[9,11]
Ideal observer	$\mathrm{SNR}_{\mathrm{i}}^2 = \displaystyle\iint \lvert G(\nu)\rvert^2\,\frac{\mathrm{MTF}^2(\nu)}{\mathrm{NPS}(\nu)}\,\mathrm{d}\nu_x\,\mathrm{d}\nu_y$	[12–14]

The **Strehl ratio** is the oldest of the four, introduced by Strehl in 1894 for telescope optics [10] and standard practice in astronomical

and lithographic optics since. Wolf et al. [5] appear to be the first to correlate it with the performance of a perception network. Being linear in the MTF it reports only how much energy is left in the core of the point spread function (PSF), and independent of form and orientation of the PSF.

The **optical informative gain** was introduced by Wolf et al. [5] to remove exactly that limitation. Because the MTF is the characteristic function of the intensity distribution in the image plane, it determines that distribution's moments, and squaring the MTF makes the metric sensitive to the second moment — to the width and orientation of the point spread function, not only to the energy in its core. Two lenses whose point spread functions differ in shape but carry the same core energy receive the same Strehl ratio and different OIG values.

The **information capacity** treats the camera as a communication channel [9,11], an argument going back to photographic science of the 1960s and revived for digital cameras in the 2000s. On a slanted edge of peak-to-peak amplitude $V_{P\text{-}P}$ the signal power follows from the MTF as $V_{P\text{-}P}^2 \mathrm{MTF}^2(f)/12$, which is what puts the MTF inside the logarithm. The kernel $\mathrm{MTF}^2/\mathrm{NPS}$ is, up to a factor, the noise-equivalent quanta of the medical-imaging literature [12], and it is invariant under uniform linear filtering, since sharpening and low-pass filtering scale numerator and denominator alike. Two measurement routes exist: the slanted edge, from which NPS follows by inverting the ISO 12233 binning [9], and a single dead-leaves target, from which Jenkin and Artmann [15] obtain MTF and NPS together and validate the result both in the RAW domain and after the image signal processor. For a test station with limited chart area and a fixed tact time, that difference may decide which metric is adoptable at all.

The **ideal-observer signal-to-noise ratio** asks how detectable a specified object is rather than how much information the channel carries. It originates in signal detection theory as applied to medical imaging, where it is codified in ICRU Report 54 [12] as the detectability index d', and was brought into automotive imaging by Kane [13] and Jenkin [14]. It is the same kernel as for Informationa Capacity, weighted by the spectrum of the object to be detected, so it is not a single number but a curve over object size; reducing it to a scalar requires a committed target size and contrast, which is a statement about the driving task

rather than about the camera. Two properties matter for production use. It is affected by the image signal processor, unlike the capacity, so it qualifies the camera rather than the lens. And for a unity MTF and pure shot noise it reduces to the Rose model [14], whose Rose criterion of $\mathrm{SNR_i} \gtrsim 5$ for reliable detection is one of only two absolute criteria available anywhere in this family — the other being the saturation threshold of Sec. 4. Every other metric considered here yields a comparison between systems, not a limit.

3.2 Properties and requirements

Tab. 2 summarises the properties and requirements of the four metrics. It is intended to be read along three axes, each of which constrains where in the supply chain a metric can be applied. The first is what has to be present to measure it: the Strehl ratio and the OIG require no noise information and can therefore be evaluated on a lens before a sensor is attached, or on a simulated wavefront during design. Measured instead on a complete camera they become sensor-dependent as well: the measured SFR already contains the pixel-aperture MTF and the integral is truncated at the Nyquist frequency, so the result is then a property of optics and sensor together. The information capacity and $\mathrm{SNR_i}$ both need a noise power spectrum (NPS) to form a frequency-dependent SNR, and therefore a complete camera; $\mathrm{SNR_i}$ additionally responds to the image signal processor (ISP).

The second is whether a reference is needed: the Strehl ratio and the OIG are normalised to the analytic diffraction limit, while the information capacity and $\mathrm{SNR_i}$ are absolute, in bits per pixel and in standard deviations respectively. The third is whether a detection task has to be committed to, which only $\mathrm{SNR_i}$ requires, through the object spectrum G; that commitment is the price of the one metric in the group carrying an absolute detection criterion.

4 Correlating metrics and CV performance

Good values of an optical quality metric are not the goal; robust CV performance is, because that is what a reliable driver assistance function rests on. The work on integral metrics arose from the gap de-

Table 2: The four integral metrics, with MTF50 as the scalar they are meant to displace. Each of them integrates the transfer function over spatial frequency; they differ in the integrand, in whether a diffraction-limited reference is required, and in what has to be measured in addition. $K(f) = \mathrm{MTF}^2(f)/\mathrm{NPS}(f)$ is the noise-weighted kernel, μ the mean linear signal and G the spectrum of the object to be detected; the definitions are collected in Tab. 1.

Metric	Integrand	Reference	Also needs	Measured on
MTF50	one point	–	–	optics or camera
Strehl ratio	MTF	diffr. limit	–	optics or camera
OIG	MTF^2	diffr. limit	–	optics or camera
Capacity C	$\log_2(1+\mu^2 K)$	–	NPS	camera
$\mathrm{SNR}_\mathrm{i}^2$	$\lvert G\rvert^2 K$	–	NPS, task	camera + ISP

scribed in Sec. 2, and this section reviews the evidence that the new metrics partly close this gap.

4.1 Windshield Aberrations and Semantic Segmentation

Wolf et al. [5] examined how the wavefront aberrations introduced by a car windshield affect a perception network. They measured the wavefront of a sample windshield with a Shack–Hartmann sensor, parameterised it by second-order Zernike coefficients, and propagated it through a Fourier-optics model to degrade KITTI images; two segmentation networks were evaluated by mean intersection over union and mean expected calibration error. Four optical merit functions were computed for each windshield state: refractive power, MTF at half Nyquist, Strehl ratio and OIG.

The current measurement standard for windscreens — refractive power — does not correlate with CV performance at all. The MTF at half Nyquist exhibits a loose threshold function with high spread. Contrarily, the Strehl ratio and the OIG exhibit a monotonic functional relationship to both benchmark metrics within the stated uncertainty intervals.

4.2 Information Capacity and Object Detection

Geever et al. [6] tested the information capacity against object detection under controlled conditions, using a synthetic target that combines a

slanted edge with rendered pedestrians, a car and a cyclist at a range of sizes, so that image quality and detection performance are measured on the same image. Quality was varied through two contrast and two Gaussian blur parameters, giving 1400 combinations at a constant Poisson shot noise level, and four detectors pre-trained on COCO were evaluated: YOLOv10m, YOLOv10n, Faster R-CNN and RT-DETR.

The information capacity correlates with detection accuracy more strongly than MTF50 does on every measure the authors report. The Spearman rank correlation lies between 0.89 and 0.98 across the four detectors, against 0.65 for MTF50. The relation saturates at a knee beyond which further capacity yields little additional accuracy, and the authors suggest that such a knee could serve as a threshold. Introducing a structured background shifts the curve towards higher capacity, that is, a more complex scene requires a better camera for the same detection performance. Imatest is carrying these metrics into ISO/TC42 as the draft standard ISO 23654 [16].

4.3 Ideal-Observer SNR and Matched-Filter Detection

For $\mathrm{SNR_i}$ the evidence has a different character. The metric is derived from a detection task, so its natural reference is a detector that is itself optimal, and two studies provide that.

Skorka and Kane [17] captured a 1951 USAF resolution target with three CMOS image sensors of different pixel size across a range of illumination levels, and compared the normalised cross-correlation of a matched filter on the tribar patterns with $\mathrm{SNR_i}$ computed from the measured MTF and noise power spectrum. Both methods rank the three sensors alike, the 2.2 µm pixel performing best except at high light levels and large pattern size, where the larger pixel catches up. The correspondence is closer than a ranking: $\mathrm{SNR_i}$ crosses the value 5 at approximately the brightness at which the measured cross-correlation reaches 0.95 for the two largest patterns, and that transition moves to higher illumination as the pattern shrinks, so a smaller object requires a higher signal-to-noise ratio for the same probability of detection. Jenkin [14] obtains the same result within a simulated imaging model: predicted $\mathrm{SNR_i}$ closely follows measured matched-filter performance.

What is missing is the step from an optimal detector to a learned one:

no study appears to have correlated SNR_i with the performance of a neural network on automotive data, and outside automotive imaging a CNN has been shown to reach ideal-observer sensitivity for some classes of signal but to fall well below it for others [18,19].

This leaves SNR_i in a particular position. It is the best founded of the four metrics, being the only one derived from a detection task and the only one carrying an absolute criterion, and at the same time the least validated against the kind of observer that automotive perception actually uses.

5 Discussion and Recommendation

Returning to the requirements of Sec. 2, all four metrics are scalars, or reduce to scalars once a task is committed to, so the first requirement is met. Measurability within an EOL cycle is met for the Strehl ratio and the OIG, which need only an SFR measurement, and is plausible but not demonstrated for the two noise-dependent metrics; the dead-leaves route of Jenkin and Artmann [15] is relevant here precisely because it yields MTF and noise power spectrum from a single target. Monotonicity across the tolerance window is the requirement that none of them yet satisfies completely: the Strehl ratio and the OIG are arguably monotonic, while Geever et al. observe a saturation above which the metric is no longer informative about performance. Neither study varies the optics as a production population does, with field-dependent and asymmetric point spread functions rather than isotropic blur. That is a gap for the industry to close on its concrete product.

That motivates the recommendation this contribution puts to the community, and in particular to those writing EOL specifications. Cameras should be qualified with at least one integral metric alongside MTF/SFR now, before any standard is finished and without initially attaching a limit to it. The additional effort is modest — for the Strehl ratio and the OIG it is a further evaluation of an SFR that is already recorded — and the resulting data are the only route to a perception-aware inspection limit grounded in production reality rather than in a simulation study. The standardisation work in IEEE 2020 [1] and the ISO 23654 draft [16] will require exactly such data, and the correlations cannot be reconstructed retrospectively from parts that were

never measured. Which of the four to record depends on where in the supply chain the measurement is made, and Tab. 2 is intended as a guide to that choice rather than as a ranking. The argument here is not that one of these metrics is perfect yet, but that the scalar currently in use has been shown not to be, and that the data needed to settle the question have to start being collected.

References

1. IEEE, "IEEE standard for automotive system image quality," IEEE Std 2020-2024, 2024, approved 11 December 2024, published 21 March 2025.
2. A. Braun, "Automotive mass production of camera systems: Linking image quality to AI performance," *tm – Technisches Messen*, vol. 90, no. 3, pp. 205–218, 2023.
3. P. Müller and A. Braun, "Simulating optical properties to access novel metrological parameter ranges and the impact of different model approximations," in *Proc. IEEE Int. Workshop on Metrology for Automotive (MetroAutomotive)*, 2022, pp. 133–138.
4. P. Müller and A. Braun, "MTF as a performance indicator for AI algorithms?" in *Proc. IS&T Int. Symp. on Electronic Imaging: Autonomous Vehicles and Machines 2023*, vol. 35, 2023, pp. 125-1–125-7.
5. D. W. Wolf, M. Ulrich, and N. Kapoor, "Sensitivity analysis of AI-based algorithms for autonomous driving on optical wavefront aberrations induced by the windshield," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV) Workshops — BRAVO*, 2023, pp. 4102–4111.
6. D. Geever, T. Brophy, D. Molloy, E. Ward, R. George, N. Koren, M. Glavin, E. Jones, and B. Deegan, "Information capacity as a predictor of perception performance," *IEEE Open Journal of Vehicular Technology*, vol. 7, pp. 708–722, 2026.
7. IATF, "Quality management system requirements for automotive production and relevant service part organizations," IATF 16949:2016, 1st ed., 2016.
8. P. Müller, A. Braun, and M. Keuper, "Examining the impact of optical aberrations to image classification and object detection models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 3, pp. 2139–2153, 2026.

9. N. L. Koren, "Image information metrics from slanted edges: A toolkit of metrics to aid object recognition, machine vision, and artificial intelligence systems," in *Proc. IS&T Int. Symp. on Electronic Imaging: Image Quality and System Performance XXI*, vol. 36, 2024, pp. 256-1–256-17.
10. K. Strehl, *Theorie des Fernrohrs auf Grund der Beugung des Lichts*. Leipzig: J. A. Barth, 1894.
11. N. L. Koren, "Measuring camera information capacity with slanted edges," in *Proc. IS&T Int. Symp. on Electronic Imaging: Image Quality and System Performance XX*, vol. 35, 2023, pp. 454-1–454-9.
12. ICRU, "Medical imaging — the assessment of image quality," International Commission on Radiation Units and Measurements, Bethesda, MD, Tech. Rep. ICRU Report 54, 1995.
13. P. J. Kane, "Signal detection theory and automotive imaging," in *Proc. IS&T Int. Symp. on Electronic Imaging: Autonomous Vehicles and Machines Conference 2019*, vol. 31, 2019, pp. 027-1–027-8.
14. R. B. Jenkin, "Comparison of detectability index and contrast detection probability," *Journal of Imaging Science and Technology*, vol. 63, no. 6, pp. 060 405-1–060 405-9, 2019.
15. R. Jenkin and U. Artmann, "Measurement of noise equivalent quanta (NEQ) using the dead leaves technique," in *Proc. IS&T Int. Symp. on Electronic Imaging: Autonomous Vehicles and Machines 2024*, vol. 36, 2024, pp. 108-1–108-7.
16. ISO/TC 42, "Digital imaging — image information metrics," ISO 23654, under development (ISO/TC 42/WG 18), 2026. [Online]. Available: https://www.iso.org/standard/87530.html
17. O. Skorka and P. J. Kane, "Object detection using an ideal observer model," in *Proc. IS&T Int. Symp. on Electronic Imaging: Autonomous Vehicles and Machines 2020*, vol. 32, 2020, pp. 041-1–041-7.
18. F. H. Reith and B. A. Wandell, "A convolutional neural network reaches optimal sensitivity for detecting some, but not all, patterns," *IEEE Access*, vol. 8, pp. 213 522–213 530, 2020.
19. K. Omer, L. Caucci, and M. Kupinski, "Limitations of CNNs for approximating the ideal observer despite quantity of training data or depth of network," *Journal of Imaging Science and Technology*, vol. 64, no. 6, pp. 060 408-1–060 408-11, 2020.

Non-contact brain-computer interface based on pupillary light reflex with extended frequency range

Niels-Ole Rohweder, Mohammadrasoul Alizadeh, and Christian Rembe

Clausthal University of Technology, Institute for Electrical Information Technology,
Leibnizstraße 28, 38678 Clausthal-Zellerfeld

Abstract Brain-computer interfaces (BCI) allow non-manual control of robots and machines, but typically derive triggers from brain activity using intrusive EEG-electrode setups. Here, we present a proof-of-concept for a fully remote, non-contact BCI based on the pupillary light reflex (PLR). We use a camera to record pupillary oscillations, stimulated by flashing lights, to trigger pick-and-place tasks on a cobot. Flashing frequencies of 2.5 Hz and 3.0 Hz extend the previously used frequency range for PLR-BCI. In an experiment with four participants, we achieved a classification accuracy of 91.5% and ITR of 3.6 bits/min. Post-hoc analysis and variation of trigger thresholds resulted in per-participant accuracy of 100 % and ITR of 6.4 bits/min. Reducing post-FFT smoothing did not substantially reduce latency when thresholds were re-optimised for maximum accuracy, indicating a more fundamental latency–accuracy tradeoff beyond effects of post-FFT smoothing.

Keywords Brain-computer interface, pupillary light reflex, pupillometry, human–robot interaction, cobot

1 Introduction

An important type of interface between humans and robots, or machines in general, is the brain-computer interface (BCI). As opposed

to physically controlling the computer, e.g. using hands, this particular type of interface doesn't require active movements by the user. As such, it can be used for people with severe motor disabilities, e.g. amyotrophic lateral sclerosis (ALS) or spinal cord injuries [1, 2]. Inputs can be used to control wheelchairs or even to communicate. Most BCIs are EEG-based and exploit distinct neurological brain-response patterns that correspond to certain, sometimes induced, states [3, 4]. An effect often used here is the Steady-State Visual Evoked Potential (SSVEP), in which brain activity (i.e., the band power) in a certain frequency band is read out [5–7]. When visually perceiving, for instance, a flashing light stimulus of a fixed frequency, the brain activity at that particular frequency band increases, for as long as the stimulus persists. By simultaneously presenting stimuli of differing frequencies, the user can focus their gaze and attention on a single stimulus, and thus trigger responses (e.g., wheelchair movements) that are mapped to the stimuli frequencies. In the case of EEG-based BCI, the read-out of the brain activity will be performed using a full EEG electrode setup on the head, which is complicated and uncomfortable for the user. This method, while reliable, is thus problematic in real-world settings. Ideally, a remote, non-contact method of read-out should be used. To that end, Sato and Nakatani recently proposed a read-out exploiting the pupillary light reflex (PLR) of the human eye [8, 9]. Within a certain frequency range, the pupil will follow the flashing light, contracting and expanding in response to the light. These pupillary oscillations can be recorded by cameras and analysed, and the so-obtained peaks in the frequency spectrum can be used to trigger desired responses. In their paper, Sato and Nakatani showed that such a setup using six different flashing frequencies shown on a head-mounted display could achieve similar results to full SSVEP-setups. Sato and Nakatani supposed a biological limit of the pupillary oscillation of 2.5 Hz, for which reason they limited their blinking frequencies in a range between 0.9 and 1.25 Hz. Similarly, Jiang et al. compared EEG signal-to-noise ratio with the one obtained by measuring pupil size, concluding that a low frequency range below 3 Hz should be chosen for PLR-based BCIs [10]. At 3 Hz, their classification accuracy dropped below 80%, even using only a single stimulus frequency at a time, not an actual BCI setup.

However, while it is true that above 1–2 Hz the frequency response of the human pupil declines with approximately 60 dB/decade (a drop by

a factor of 8 for doubling the frequency), and thus the amplitudes of the oscillations rapidly decrease, this is more a question of the precision of the measurement setup, rather than a hard biological limit. As we have shown, for instance, using our self-developed, extremely precise pupillometer, it is possible to obtain signal-to-noise ratios well above 30 dB (i.e., a factor of 1000) for oscillations at 4.2 Hz [11–13]. The oscillation amplitudes at these frequencies are on the order of a few micrometres. Increasing the upper limit of the useful frequency range for PLR-based BCIs can increase the quantity of information that can be transmitted. Therefore, we present in this paper a PLR-setup similar to Sato and Nakatani's, but using two blinking frequencies of 2.5 Hz and 3.0 Hz, thereby improving on the state of the art. To showcase the setup, we use it to control a robotic arm performing a pick-and-place task. Our proof-of-concept can be easily adapted to other tasks. Additionally, we perform a post-hoc analysis of the recorded data to evaluate the influence of spectral weighting, trigger thresholds, and post-FFT temporal smoothing on classification accuracy and response latency.

2 Methods

2.1 Cobot experiment

Setup

The experimental BCI setup consists of two units. Fig. 1 shows the pupillometer and its schematic top-down view. Fig. 2 shows the cobot. Both are linked via their computer workstations using a TCP bridge, which transmits measurements every 0.5 s from the pupillometer unit. The cobot unit evaluates the measurements and triggers cobot action. The participant sits in front of the pupillometer, while the pupil of the right eye is illuminated with an IR floodlight (6 W, 850 nm) and recorded by the pupillometer. The pupillometer's camera-lens system consists of a telecentric lens with a filter blocking visible light and a 12.4 MP camera (ids U3-31L0CP Rev. 2) with a field-of-view of 2500 $\times$ 2500 pixels, running at 32 fps. Further technical details can be found in [13]. The stimuli are created by two green LED arrays of 3.5$\times$2 cm, placed at a distance of 14 cm in front of the eye, with a lateral distance of 4 cm between the two stimuli. The arrays are driven by a frequency

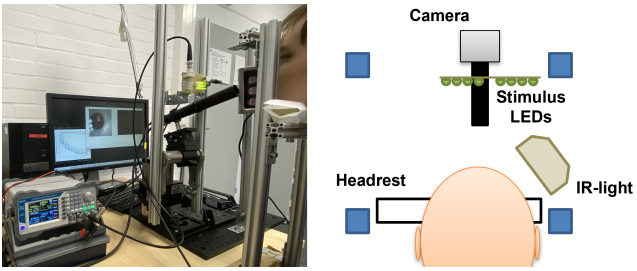


Figure 1: Left: The pupillometry setup as used in the experiment, including camera and lens, stimulus LEDs, IR light and function generator. Right: A schematic top-down view of the pupillometer.

generator using a square-wave signal at frequencies of 2.5 and 3 Hz, respectively. The cobot, a 7-DOF Franka Emika "Panda", is set up at a table with access to a slide with disks and a box to place them in. Fixed trajectories between the two positions were programmed using superposed sines. Additionally, an intermediate "neutral" position was programmed, which the cobot moved to in case of specific mistriggers.

Four participants with ages ranging from 19 to 38 were recruited and written consent obtained. They were made familiar with the setup, and then prompted to steer the robotic arm to place 10 disks into the box by switching gaze between the two stimuli. The stimuli were blinking throughout the entire time of the experiment, and a trial lasted until all disks were in the box, taking between 5–10 minutes. The trials were recorded with a webcam for later analysis. This was to determine whether a triggered robot movement was intended by the user, as well as measure the time in-between user intention and triggered action.

Algorithm, cobot control and parameters

The luminance-based algorithm of [13] and [14] was used to determine the pupil area and pupil diameter, respectively, using the contrast of the dark pupil with the bright cornea, illuminated by the IR light. The so-obtained time series of pupil diameters is continuously Fourier-transformed using a sliding rectangle window of 8 s length. A full spectrum is submitted to the robotic control nominally every 0.5 s, but the received spacing is non-uniform. Two ring buffers keep the

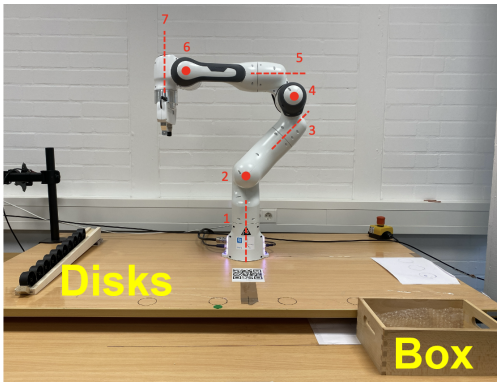


Figure 2: The 7-DOF robot ("Panda", Franka Emika) used to perform the pick-and-place task: Controlled by the PLR-BCI, the user triggers the actions of (1) picking a disk off the slide, (2) placing it in a box.

last $l = 4$ values of the 2.5 and 3 Hz frequency bins, $S_t^{2.5\,Hz}$ and $S_t^{3.0\,Hz}$, which are, respectively, mapped to a "pick" and a "place" motion of the cobot. To smooth the response and reduce spurious signals, the values in the ring buffer are averaged, creating a running mean. To account for the 2.5 Hz peak being, on average, higher than the 3.0 Hz one, due to the decreasing frequency response of the eye, the running mean of the 2.5 Hz buffer is multiplied by a factor w of 0.7,

$$S_{pick} = 0.7/l \cdot \sum_{t=-l}^{-1} S_t^{2.5\,Hz} \quad \text{and} \quad S_{place} = 1/l \cdot \sum_{t=-l}^{-1} S_t^{3.0\,Hz}$$

From this, a ratio R is calculated according to:

$$R = \frac{S_{pick}}{S_{pick} + S_{place}} \tag{1}$$

If R exceeds $t_{pick} = 0.62$, a "pick" trigger is generated. If the value is lower than $t_{place} = 0.38$, a "place" trigger is generated. If a "place" action is triggered while a "pick" action is still ongoing (or vice versa), this is classified as a mistrigger, and the robot moves to a "neutral" holding position, allowing the user to re-trigger the intended action. Triggers of the same action that is currently ongoing, as well as triggers

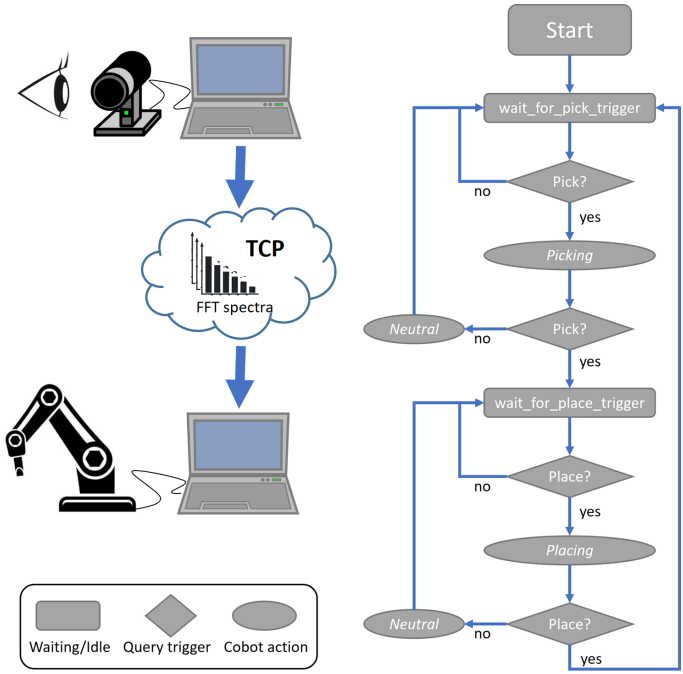


Figure 3: Left: The full BCI setup schematically. Right: State flow of the cobot experiment. Repeated triggers of the same kind are ignored, as are triggers opposite to the waiting state. Mistriggers are created when during execution of an action an opposite trigger is received. The cobot then moves to a neutral position.

identical to the current state (e.g., trigger "pick" while in pick-position) are ignored. A state flow is shown in Fig. 3. The BCI-controlled cobot thus picked up disks from a slide, and placed them in a box (Fig. 2).

The parameters appearing here – the length of the Fourier transform, the size of the ring buffer l, the normalising factor w, and the trigger thresholds t_{pick} and t_{place} – were empirically determined using a test run prior to the cobot experiment. The full spectra received by the cobot unit were saved, allowing all parameters except the length of the Fourier-window to be subsequently varied in the post-hoc analysis described below. From video data, classification accuracy and latency were extracted, from which Information Transfer Rate (ITR) according

to Wolpaw [3] was calculated:

$$\mathrm{ITR} = V\left[\log_2 N + P\log_2 P + (1-P)\log_2\left(\frac{1-P}{N-1}\right)\right]$$

where ITR is measured in bits/minute, N is the number of states (i.e., in this case N = 2), P the accuracy, and V the average selection speed in triggers/minute. The latter is calculated by multiplying the inverse mean latency by 60. Afterwards, this per-trial ITR is averaged as $\overline{\mathrm{ITR}} = (\mathrm{ITR}_1 + \mathrm{ITR}_2 + \mathrm{ITR}_3 + \mathrm{ITR}_4)/4$. A trigger was classified as intentional if it was preceded by a corresponding eye movement by at least 1 s. The limit was set empirically; in practice, a cut-off is needed to separate intentional from accidental triggers. 1 s has barely enough signal accumulated inside the 8 s-Fourier window that the SNR or the ratio R could in theory have risen to trigger cobot action. Signal-to-noise ratio (SNR) values of the 2.5 Hz and 3.0 Hz peaks were calculated from the raw spectra of the pupil diameter time series, relative to the noise floor (i.e., the average spectral density beyond 10 Hz), by averaging values whenever R was above (2.5 Hz) or below (3.0 Hz) trigger thresholds.

2.2 Post-hoc analysis

In order to verify and vary the three parameters w, l, and t_{pick} and t_{place} (normalising factor, ring buffer length and trigger thresholds) we ran a post-hoc analysis on the recorded spectral time series. For w, we calculated the mean of the spectral 2.5 Hz component and the mean of the 3 Hz component per participant and, as their ratio, $w_i = \overline{S_{i,\,place}}/\overline{S_{i,\,pick}}$, ranging from $w_R = 0.768$ to $w_E = 0.908$, with an unweighted average of $w_{avg} = 0.815$ as opposed to the value of $w = 0.7$ used in the experiment after the preliminary tests. w_{avg} was subsequently used for any analysis beyond the specific cobot experiment.

In a second step, we optimised the thresholds. The process was such that (1) all intentionally triggered cobot action was retained, (2) the narrowest neutral zone defined by $t_{pick} - t_{place}$ is obtained, while also (3) no mistriggers are added, or, where that is not possible, the amount of mistriggers is minimised. This was done for each participant individually, and for ring buffer lengths from $l = 4$ to $l = 1$, i.e. no post-FFT smoothing of the incoming spectral components.

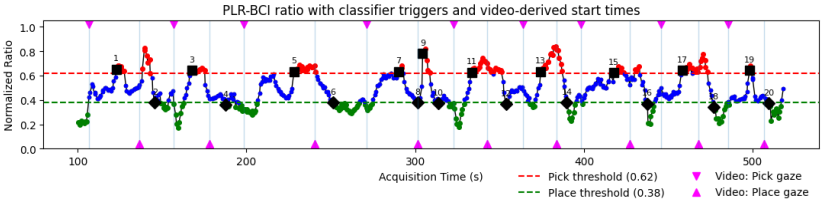


Figure 4: 2.5-to-3 Hz-ratio R over the course of the entire experiment for participant A. Thresholds of 0.62 to trigger "pick" and 0.38 to trigger "place" are marked in red and green, respectively. Actions 8 and 9 in the plot are mistriggers.

Generally, the precise numeric values for the thresholds obtained by this process are somewhat arbitrary, as they depend on the variance of the ratio R when not driven by the stimulus, which in turn is highly dependent on the participant and the generation of spectral noise (small eye and body movements, distractions, the way the gaze shifts between stimuli and so on). However, the relative changes can tell about underlying tradeoffs between quick triggering (i.e. short latency) and accurate triggering (i.e. few or no mistriggers), and additionally provide a measure for the performance of the BCI possible in theory.

3 Results and discussion

3.1 Cobot experiment

Fig. 4 shows an exemplary time course of R for participant A over the course of the experiment, including the trigger thresholds. The time course includes a mistrigger: cobot actions 8 and 9 occur without gaze shift, while the "place" gaze was only supposed to trigger cobot action 10. Fig. 5 shows exemplary raw spectra, before averaging in the ring buffer, at certain points in time for the 2.5 Hz and 3 Hz stimuli as recorded from the pupil. The SNR for the individual participants, derived from the spectral time series, are shown in Tab. 1. On average, across all 4 participants, the SNR for the 2.5 Hz stimulus was 13.2 dB, and for the 3.0 Hz stimulus, 11.1 dB. With these SNR values the setup achieved a high classification accuracy of 91.5% across all participants. Mean gaze-action latency was 12.4 s, corresponding to an ITR of 3.6 bits/minute.

Table 1: Signal-to-noise ratio.

Participant	SNR 2.5 Hz (dB)	SNR 3.0 Hz (dB)
N	15.8	13.8
E	11.1	7.5
A	12.6	11.0
R	11.9	9.8
Overall Average	13.2	11.1

3.2 Post-hoc analysis

Optimising the individual thresholds as described in section 2.2 results in 100 % accuracy for the participants and the 4-element ring buffer. The mean latency decreases to 10.415 s (Tab. 2), and ITR increases to 6.4 bits/minute. The notable result is that mean latency remains almost unchanged when the ring-buffer length is reduced, provided that the thresholds are re-optimised to maintain maximum accuracy. Reducing the amount of smoothing causes R to respond more rapidly to changes in the incoming spectra, which in isolation would lead to earlier threshold crossings. At the same time, however, the increased variance requires a wider neutral zone in order to avoid mistriggers. These two effects compensate one another, so that mean latency remains approximately constant across buffer lengths. The main effect of the ring buffer therefore is to reduce the variance of R, allowing more robust classification. Indeed, at $l = 1$, i.e. no post-FFT smoothing, all participants had either mistriggers or missed triggers, regardless of where the thresholds were placed; and so, beyond a fixed delay, the ring buffer is largely latency-neutral when optimised for maximum accuracy. This shows the more fundamental tradeoff between accuracy and speed, where the underlying data input limits performance, consistent with the imperfect separation of the "pick" and "place" triggers in the raw spectra. Often, rather than showing a single clear peak in the spectra, one was only slightly higher than the other (Fig. 5, right), even though the user intended to trigger an action. Temporal averaging – both implicitly by the 8-s-Fourier window, and the explicit smoothing in the ring buffer – extracts clear triggers, as evidenced by the high classification accuracy, but comes at the cost of increasing the total time to make

Table 2: Post-hoc classifier performance for different ring-buffer lengths l. Thresholds T and latency Δt are averaged across participants; mistriggers are summed.

l	$\overline{T}_{\text{pick}}$	$\overline{T}_{\text{place}}$	$\overline{\Delta T}$	$\overline{\Delta t}$ (s)	Mistriggers
4	0.628	0.434	0.194	10.415	0
3	0.641	0.423	0.218	10.661	3
2	0.678	0.408	0.270	10.660	8
1	0.746	0.358	0.389	10.658	25

a decision. For our purposes, this validates the use of ring buffering.

Further improvements to increase the current, comparatively low ITR therefore should mostly focus on the pupillometer unit, particularly by increasing SNR and allowing for shorter Fourier windows. This can either be done by better noise suppression, or stronger signals from the outset. Conceivably, one could improve the setup – for instance, the positioning or spatial separation of the flickering stimuli – to achieve a better separation of pupillary light influx and thus unambiguous separation of the spectral peaks. Increasing transmission rate between pupillometer and cobot and keeping it stable might also improve performance. The non-uniform spacing of the spectra received by the cobot unit, partly due to a bug in the TCP implementation, may have caused additional delay, as preliminary attempts at interpolation of the missing spectral values resulted in a decrease in latency of 10 % for one participant. Finally, an improvement might be made to the monocular setup. It is known that inducing a light stimulus in one eye

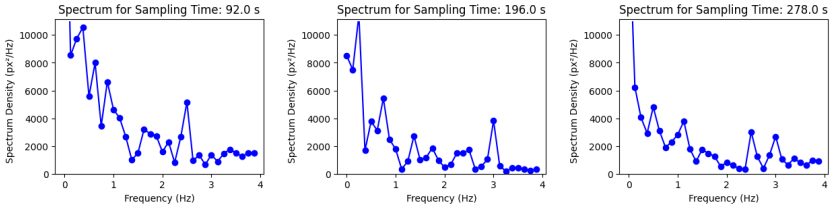


Figure 5: Raw spectra as calculated by the FFT of the pupil diameter time series. Left and centre: Raised 2.5 Hz and 3 Hz activity during pick and place times. Right: A superposition of 2.5 and 3.0 Hz oscillations, even though the participant intended to trigger one action, corresponding to an undefined trigger situation.

produces a response for both eyes, and thus both eyes should obtain the same predominant stimulus, in order to avoid superpositions of the two signals. Presenting two screens, like Sato and Nakatani did, one for each eye, both containing the full set of stimuli, might produce much clearer responses. Indeed, as Sato and Nakatani showed, splitting the stimuli across the eyes reduced classification accuracy substantially.

4 Conclusion

In this paper, we showed a proof-of-concept for a pupillary light reflex (PLR)-based brain-computer interface (BCI), using the frequency range of 2.5 Hz and higher, which has not been previously presented. Flickering stimuli of 2.5 Hz and 3.0 Hz expand the bandwidth available for such non-contact BCI. Combining our self-developed pupillometer, which has a spatial precision on the order of a few hundred nanometres, with a Franka "Panda" cobot, we created a demo to showcase the concept. Measured pupil size was Fourier-transformed in real time and the 2.5 Hz and 3.0 Hz spectral components were used to trigger "pick" and "place" movements. Four participants were then recruited to place 10 disks in a box, using this demo setup. We achieved a classification accuracy of 91.5% and an ITR of 3.6 bits/minute. Post-hoc analysis of the recorded data showed a theoretical ITR of up to 6.4 bits/minute. Additionally, under the constraint of optimising for maximum accuracy, we found that the post-FFT smoothing in a running-mean ring buffer introduced no additional length-dependent delay, but rather remained at an approximately constant total delay, highlighting the more fundamental nature of the tradeoff between accuracy and speed.

Acknowledgements

This work was financed by the State of Lower Saxony as part of the project "KEIKO", using the SPRUNG funding grant 76251-1337/2022.

References

1. J. P. Donoghue, "Bridging the brain to the world: A perspective on neural interface systems," *Neuron*, vol. 60, no. 3, pp. 511–521, 2008.

2. L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.
3. J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, June 2002.
4. R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of eeg-based brain-computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, p. 011001, 2019.
5. D. Regan, *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. Elsevier, 1989.
6. D. Zhu, J. Bieger, G. Garcia Molina, and R. M. Aarts, "A survey of stimulation methods used in ssvep-based bcis," *Computational Intelligence and Neuroscience*, vol. 2010, pp. 1–12, 2010.
7. S. T. Morgan, J. C. Hansen, and S. A. Hillyard, "Selective attention to stimulus location modulates the steady-state visual evoked potential," *Proceedings of the National Academy of Sciences*, vol. 93, no. 10, pp. 4770–4774, 1996.
8. A. Sato and S. Nakatani, "Noncontact brain–computer interface based on steady-state pupil light reflex using independent bilateral eyes stimulation," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 3386–3390.
9. A. Sato and S. Nakatani, "Independent bilateral-eye stimulation for gaze pattern recognition based on steady-state pupil light reflex," *Journal of Neural Engineering*, vol. 19, no. 6, p. 066046, 2022.
10. L. Jiang, X. Li, Y. Wang, W. Pei, X. Gao, and H. Chen, "Comparison of pupil size and visual evoked potentials under 1–6hz visual stimulation," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020, pp. 3054–3057.
11. N.-O. Rohweder, C. Rembe, and J. Gertheiss, "Towards a remote eeg for use in robotic sensors," in *Forum Bildverarbeitung 2020*. KIT Scientific Publishing, October 2020, pp. 225–237.
12. N.-O. Rohweder, J. Gertheiss, and C. Rembe, "Sub-micron pupillometry for optical eeg measurements," *tm – Technisches Messen*, vol. 88, no. 7-8, pp. 473–480, August 2021.
13. N.-O. Rohweder and C. Rembe, "High precision pupillometry," *tm – Technisches Messen*, vol. 90, no. S1, pp. 114–119, 2023.
14. N.-O. Rohweder and C. Rembe, "Pupillary spectra for use in optical electroencephalography," in *Applied Optical Metrology VI*, ser. Proceedings of SPIE, vol. 13603, September 2025, p. 136030D.

Laser-tracker-assisted camera-to-robot calibration for mobile robots

Jan Andre Rudolph, Öykü Kandemir, and Markus Ulrich

Institute of Photogrammetry and Remote Sensing (IPF),
Karlsruhe Institute of Technology (KIT),
Englerstr. 7, 76131 Karlsruhe, Germany

Abstract We present a laser-tracker-assisted hand–eye calibration method for camera-equipped mobile robots. The method combines laser-tracker-based 3D metrology with camera-based 2D observations. Building on our previous laser-tracker-assisted camera-to-robot calibration method for ground-observing mobile robots, we present a generalized formulation for calibrating the camera pose in the coordinate system of tracker-localized mobile robots. The new approach relaxes assumptions of our previous method on robot and camera configuration by chaining multiple calibration targets resulting in a more general approach supporting various camera-equipped mobile robot systems.

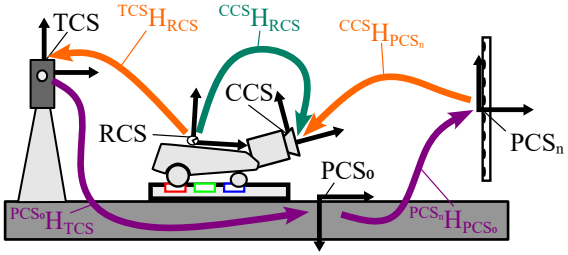


Figure 1: Coordinate systems and transformations relevant for camera-to-robot calibration, extending [1]: tracker (TCS), robot (RCS) and calibration plates (PCS_i). The robot-mounted camera (CCS) observes PCS_n. The reference plate (PCS_0) can be absent in the calibration step. Purple denotes known transformations, orange measurements, and green the result ${}^{RCS}H_{CCS}$.

1 Motivation

This work is based on the mobile robot RITA [1], designed for high-accuracy positioning and measurement tasks. RITA is intended for operation in large indoor industrial facilities, where it navigates on planar factory floors while performing positioning and measurement tasks. A camera integrated in the mobile robot supports accurate 3D reconstruction and localization, requiring reliable and accurate calibration of its pose with respect to the robot coordinate system. For this, we combine external 3D metrology from a laser tracker with camera-based target observations to solve the camera-to-robot calibration problem for tracker-localized mobile robots. This setting is challenging because the camera coordinate system is not directly observable by the laser tracker, and the motion of mobile robots often does not provide the pose diversity required by conventional hand–eye calibration approaches.

Hand–eye calibration is well established for vision-guided industrial robots. Typically, these approaches assume sufficiently diverse calibration motions to estimate the rigid camera-to-robot transformation (i.e., the hand–eye pose) unambiguously and accurately [2–6]. In contrast, mobile robots are often restricted in their motion, e.g., in the case of RITA an approximately planar motion. This makes the direct application of conventional hand–eye calibration methods difficult, since the available robot poses do not sufficiently constrain all components of the camera-to-robot transformation. A similar challenge occurs in the hand-eye calibration of SCARA robots, whose rotational joints are all parallel. In this case, the hand-eye pose is only observable up to an unknown translation along the rotation axes. Although this ambiguity can be resolved by moving the tool to a predefined height [7], this approach is not applicable to the mobile robot, as it would require vertical motion and physical access to the center of the robot coordinate system. The present work addresses this gap through a metrology-supported target-registration workflow.

This work builds on our previous laser-tracker-assisted camera-to-robot calibration method for ground-observing mobile robots [1]. The method utilizes a dedicated referencing plate that combines reflector nests for laser-tracker-based pose estimation with an integrated camera calibration target. The present work generalizes this plate-based approach beyond the ground-observing setup. Instead of relying on

an in-house-designed drive-on referencing plate, the proposed method successively links multiple calibration targets. The chain removes the need for a dedicated combined plate; we nevertheless reuse the existing metrologically characterized reference plate as a convenient tracker anchor. Previous work on hand–eye calibration and calibration uncertainty from both machine-vision and geodetic perspectives [8–10] is also relevant in this context, especially [7], which dealt with restricted pose diversity, exploiting a SCARA-specific geometric constraint. In contrast, we focus on a practical calibration procedure for camera-equipped mobile robots localized by a laser tracker rather than on manipulator kinematics or SCARA-like configurations.

The contribution of this work is a generalized laser-tracker-assisted camera-to-robot calibration workflow that extends our previous work [1], which only supported a nadir-camera setup, to more general camera mounting configurations. The method combines the laser tracker as an external metric reference with image-based observations of calibration targets. It introduces a chained target-registration procedure that transfers the spatial relation from the camera-observed calibration target to the reflector nests. We evaluate the approach on a real mobile robot by assessing the precision of the estimated camera-to-robot transformation and determining its absolute accuracy through a 3D measurement task. The results are compared against ground-truth measurements obtained manually using a laser tracker.

2 Method

Our proposed method generalizes our previous work [1], which relied on a dedicated reference plate for calibration that incorporates reflector nests, which support the reflectors of the laser tracker, and a camera calibration target on a common plane. The nadir camera directly observed the reference plate. Instead, our new approach does not require a nadir camera but also works with cameras in arbitrary orientations (for an example of a robot with a front-looking camera see Fig. 1). This is achieved by distinguishing between the reference plate and the plate that is observed by the camera yielding one calibration object. This allows to place the observed plate such that it can be acquired with the robot's camera.

Let ${}^{A}H_{B}$ be the homogeneous rigid transformation that maps 3D points from coordinate system B to A, and ${}^{A}P_{j}$ a homogeneous 3D point j expressed in A. We distinguish the tracker coordinate system (TCS), robot coordinate system (RCS), coordinate system of the robot-mounted camera (CCS), auxiliary-camera coordinate system in image i (ACS_i), and coordinate system of calibration plate i (PCS_i) (see Fig. 1).

2.1 Calibration Object

To create the calibration object, the relative pose of the observed plate with respect to the reference plate must be determined prior to hand-eye calibration by taking an image showing both plates with a calibrated auxiliary camera. If both plates cannot be imaged in a single image, alternatively, a chain of calibration plates can be linked by multiple images (see Fig. 2). Adjacent images in this chain must share a common fixed target. Note that because these images contain two calibration targets, both calibration targets must be distinguished and identified correctly during the automatic detection.

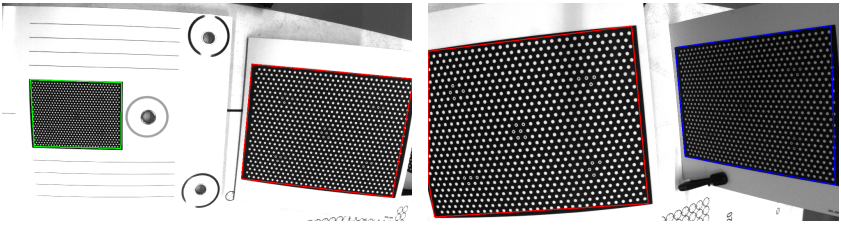


Figure 2: Two successive auxiliary camera images with detected HALCON [11] calibration targets: PCS_0 (green) and PCS_1 (red) in image 1; PCS_1 (red) and PCS_2 (blue) in image 2. The shared target links the observations. Images are cropped for visualization.

An auxiliary camera acquires n images from different poses with coordinate systems ACS_i, $i = 1, \ldots, n$. Image i shows the adjacent calibration plates PCS_{i-1} and PCS_i. Combining the two target poses detected in image i yields ${}^{\mathrm{PCS}_i}H_{\mathrm{PCS}_{i-1}} = ({}^{\mathrm{ACS}_i}H_{\mathrm{PCS}_i})^{-1}({}^{\mathrm{ACS}_i}H_{\mathrm{PCS}_{i-1}})$. The complete image chain connects the reference plate PCS_0 to the plate PCS_n that is observed by the robot during calibration: ${}^{\mathrm{PCS}_n}H_{\mathrm{PCS}_0} = \overleftarrow{\prod}_{i=1}^{n} {}^{\mathrm{PCS}_i}H_{\mathrm{PCS}_{i-1}}$ (the arrow indicates multiplications from the left).

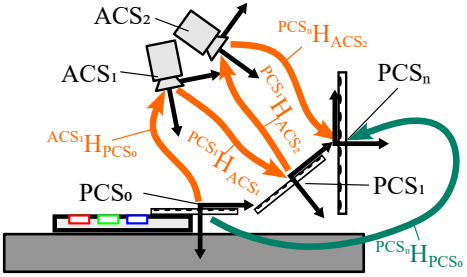


Figure 3: Example target chain with $n = 2$ auxiliary camera poses. PCS_0 is the reference plate, PCS_1 the intermediate plate, and $\mathrm{PCS}_n = \mathrm{PCS}_2$ the plate observed by the robot. Auxiliary camera 1 observes PCS_0 and PCS_1, auxiliary camera 2 PCS_1 and PCS_2. Observed poses are displayed in orange, the resulting pose in green.

As described in [1] in more detail, a stereo measurement yields the coordinates of the 3D centers of three circular reflector nests ${}^{\mathrm{PCS}_0}P_j$, $j \in \{r, g, b\}$ (red, green, blue) relative to the reference calibration plate, while also incorporating an offset of the distance to the centers of the inserted reflectors. The centers of the three reflectors ${}^{\mathrm{TCS}}P_j$ are measured by the laser tracker and used together with ${}^{\mathrm{PCS}_0}P_j$ to estimate ${}^{\mathrm{PCS}_0}H_{\mathrm{TCS}}$ [1, Eq. (13)].

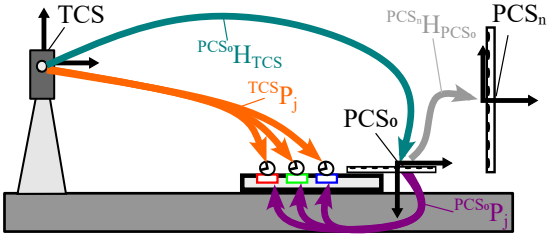


Figure 4: Reference-plate registration [1, Eq. (13)]. The three laser tracker measured reflector centers ${}^{\mathrm{TCS}}P_j$ (orange) are matched to the known ${}^{\mathrm{PCS}_0}P_j$ (purple), yielding ${}^{\mathrm{PCS}_0}H_{\mathrm{TCS}}$ (green). Note that the robot-observed plate (PCS_n) is not registered directly against the laser tracker measurements.

This process yields the calibration object, which can be used to calibrate a series of robots.

2.2 Camera–Robot Calibration

To calibrate the robot, the robot-mounted camera acquires an image of the calibration target (PCS_n), which yields ${}^{\mathrm{CCS}}H_{\mathrm{PCS}_n}$, and hence, ${}^{\mathrm{CCS}}H_{\mathrm{PCS}_0} = {}^{\mathrm{CCS}}H_{\mathrm{PCS}_n}\,{}^{\mathrm{PCS}_n}H_{\mathrm{PCS}_0}$. The robot pose ${}^{\mathrm{TCS}}H_{\mathrm{RCS}}$ is obtained from the laser tracker, which measures the position of the reflector mounted on the robot, and from inertial measurement information. Alternatively, the rotation part of ${}^{\mathrm{TCS}}H_{\mathrm{RCS}}$ is obtained by the method described in [1, Eqs. (14)–(21)]. Finally, the camera-to-robot pose is composed as

$$ {}^{\mathrm{CCS}}H_{\mathrm{RCS}} = {}^{\mathrm{CCS}}H_{\mathrm{PCS}_0}\,{}^{\mathrm{PCS}_0}H_{\mathrm{TCS}}\,{}^{\mathrm{TCS}}H_{\mathrm{RCS}}. \tag{1} $$

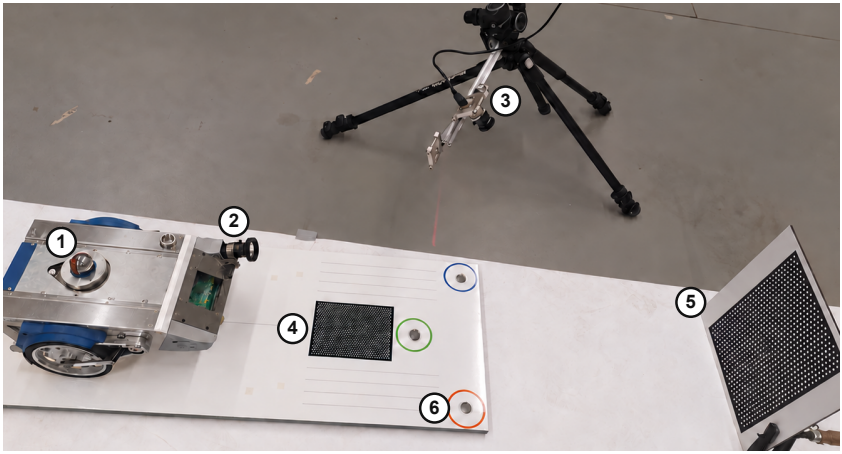
3 Results

In a first experiment, we evaluate the precision of the camera-to-robot pose, in the second experiment, its absolute accuracy. In our previous work with the nadir setup [1], repeated measurements of a printed floor mark were used for verification. Consequently, only the precision could be evaluated. Because the outward-facing camera observes a larger workspace and permits measuring reference points with a laser tracker, now we are able to evaluate the absolute accuracy as well. Fig. 5 shows the actual hardware that was used and its specifications.

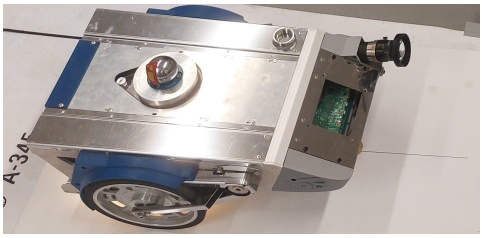
3.1 Precision

To evaluate the precision of our approach, we perform the creation of the calibration object (Sect. 2.1) and the camera–robot calibration (Sect. 2.2) eight times. In the first $N = 4$ experiments (Scene A), we used a single intermediate calibration plate (i.e., $n = 2$) and varied the poses of the intermediate and the observed plate. We chose a distance between reference and observed plate of about 0.9 m. In the second $N = 4$ experiments (Scene B) we rotated the reference plate by approx. 180 degrees about the vertical axis and used two intermediate calibration plates (i.e., $n = 3$) to cover the larger distance between reference and observed plate of about 1.3 m. Consequently, Scene A uses two chain images, whereas Scene B uses three to span the larger distance.

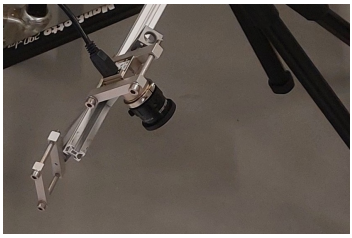
Table 1 and Figure 6 show the results of the precision evaluation for Scene A and B. The lower precision of the results of Scene B are obvious



(a)



(b)



(c)

Figure 5: Experimental setup. (a) and (b): the mobile robot guided by a Leica Absolute Tracker AT901 [12], with a Leica/Hexagon 1.5-inch break-resistant corner-cube reflector (1), and the robot-mounted high-resolution imaging system (2): IDS camera (5328 $\times$ 4608 pixels, 2.74 $\bar{}$m pixel pitch) [13] with a Schneider-Kreuznach 12 mm lens [14]. (a) and (c): the auxiliary camera system (3): IDS camera (5136 $\times$ 3856 pixels, 1.40 $\bar{}$m pixel pitch) [15] with a Schneider-Kreuznach 4.8 mm lens [16]. Furthermore, (a) shows the HALCON [11] calibration targets: the reference plate (PCS_0) is a 160 mm target (4), the robot-observed plate (PCS_n) is a 320 mm target (5). Representative of the three nests (r, g, b) described in Sect. 2.1, (6) identifies the red reflector nest r. Image (a) shows an exemplarily calibration setup in Scene A.

		RMS to scene mean		Max. pair distance	
Group	N	mm	deg	mm	deg
Scene A	4	2.449	0.124	6.041	0.288
Scene B	4	5.216	0.360	12.457	0.717

Table 1: Results of precision evaluation. RMS values of the translation and rotation component of ${}^{\mathrm{CCS}}H_{\mathrm{RCS}}$ with respect to the average translation vector and average rotation matrix. Maximum differences between pairs of experiments in translation and rotation.

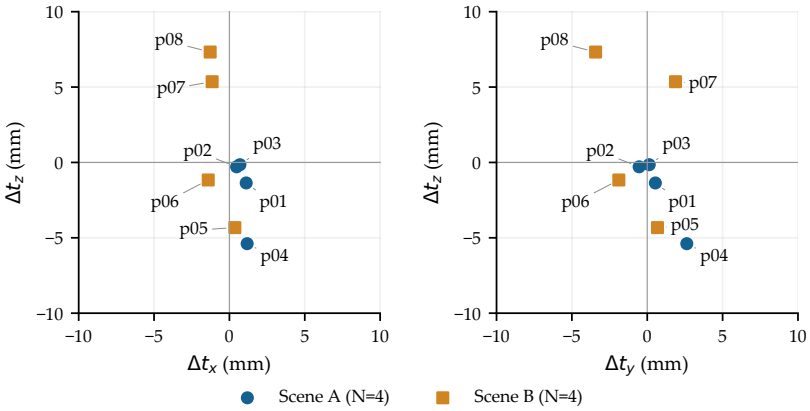


Figure 6: Results of precision evaluation. Deviation of the translation component of ${}^{\mathrm{CCS}}H_{\mathrm{RCS}}$ with respect to the mean of the eight camera-to-robot estimates. Blue circles are calibration results of Scene A, orange squares results of Scene B.

and are caused by the larger distance of the reference plate from the observed plate, which made an additional intermediate calibration target pose necessary. While for our localization application, the precision is sufficient, in general the calibration chain length should be minimized to avoid error accumulation. From Figure 6, one can see that the precision in z direction (i.e., the viewing direction of the camera) is worse than in lateral direction.

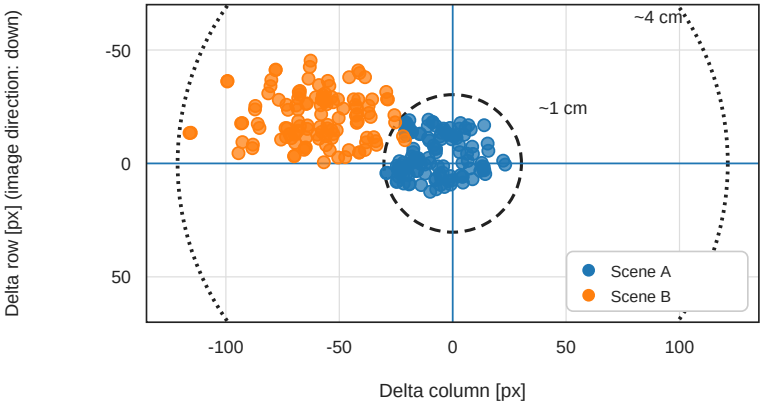


Figure 7: Results of accuracy evaluation. Offset of the measured image points to the reprojected ground truth of 240 experiments, distinguishing experiments of Scene A (blue) and B (orange). For visualization purposes, circles of points with equidistant offsets of 1 cm and 4 cm are projected into the image.

3.2 Accuracy

To evaluate the absolute accuracy of the camera–robot pose, the ground truth 3D coordinates of a marked reference point are measured with a laser tracker. Then, an image of the reference point is acquired with the robot's camera and its image coordinates are determined. This image point defines a line-of-sight in 3D space. As evaluation metrics we measure the orthogonal distance of the 3D reference point from the line-of-sight as well as its reprojection error.

To transform measurements from CCS into TCS, the robot pose is obtained from laser tracker positions and IMU orientation. Because of the planar robot motion, we assume zero roll and pitch (see [1, Eqs. (14)–(21)]). With the inverse robot pose ${}^{\mathrm{RCS}}H_{\mathrm{TCS}}$, the calibration result ${}^{\mathrm{CCS}}H_{\mathrm{RCS}}$ yields the pose ${}^{\mathrm{CCS}}H_{\mathrm{TCS}} = {}^{\mathrm{CCS}}H_{\mathrm{RCS}}\,{}^{\mathrm{RCS}}H_{\mathrm{TCS}}$ of the camera in the TCS.

The measured image point $\mathbf{u} = (row, col)^\top$ and the calibrated camera model define the line-of-sight in the CCS. Because the line-of-sight passes through the origin of the CCS it can be described by the vector ${}^{\mathrm{CCS}}\mathbf{d}$. The tracker-measured ground truth reference point ${}^{\mathrm{TCS}}P_g$ is

transformed into the CCS by ${}^{\mathrm{CCS}}\bar{P}_g = {}^{\mathrm{CCS}}H_{\mathrm{TCS}}\,{}^{\mathrm{TCS}}\bar{P}_g$ (where $\bar{P}$ denote homogeneous coordinates of P). Note that ${}^{\mathrm{CCS}}H_{\mathrm{TCS}}$ contains the calibrated camera–robot pose that is to be evaluated. Then, the orthogonal distance of the reference point to the line-of-sight is

$$e_\perp = \left\| {}^{\mathrm{CCS}}P_g - [({}^{\mathrm{CCS}}\mathbf{d})^\top {}^{\mathrm{CCS}}P_g]\,{}^{\mathrm{CCS}}\mathbf{d} \right\|. \tag{2}$$

The reprojection error is computed as

$$\Delta\mathbf{u} = \pi({}^{\mathrm{CCS}}P_g) - \mathbf{u} = (\Delta row, \Delta col)^\top, \qquad e_{px} = \|\Delta\mathbf{u}\|, \tag{3}$$

where $\pi({}^{\mathrm{CCS}}P_g)$ is the 3D reference point projected into the image by using the calibrated camera model.

For the evaluation, we moved the robot to 30 different poses and took an image of the reference point from each pose. Then, for all calibration results of the eight experiments of Scene A and B from Sect. 3.1 and for each image, we computed both evaluation metrics. Table 2 and Fig. 7 summarize the results. Note that the accuracy of all 120 results of the calibrations using Scene A is better than 1cm As already seen from the precision evaluation, the calibrations using Scene B lead to worse results. Fig. 7 clearly shows that the results of Scene A are scattered around the ground truth while those of Scene B show a significant systematic error.

Group	Metric	Mean	RMS	Max.
Scene A	$e_\perp$ (cm)	0.580	0.615	0.961
Scene A	e_{px} (px)	16.098	17.265	29.561
Scene B	$e_\perp$ (cm)	2.331	2.405	3.824
Scene B	e_{px} (px)	64.190	66.708	116.794

Table 2: Results of accuracy evaluation. Absolute errors for Scene A and B, each containing 120 experiments. The metrics $e_\perp$ and e_{px} denote the point-to-line and pixel errors.

4 Conclusion

We presented a laser-tracker-assisted camera-to-robot calibration method that extends our previous nadir-camera approach to arbitrarily oriented cameras by connecting the tracker-referenced and robot-observed calibration plates through a chain of calibration targets. The

experiments demonstrate the practical applicability of the method and also show that the chain design is critical: the precision and accuracy of all 120 Scene A evaluations are better than 1 cm, whereas the longer chain of Scene B yielded a lower precision and the accuracy results show a systematic offset. Thus, the distance between the reference and observed plates and the number of intermediate calibration targets should be minimized to limit error accumulation. Future work will isolate and quantify the individual error contributions: e.g., translation and orientation components of the robot poses, the calibration target poses, and the stability of the setup. This will help to optimize the calibration target arrangements and the methodology, and hence, to further improve accuracy and precision.

Acknowledgements

This Project is supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) on the basis of a decision by the German Bundestag.

The authors used AI tools for language refinement and generation of scripts for statistical analysis and plotting. All content was verified by the authors, who take full responsibility for the final manuscript.

References

1. J. A. Rudolph, D. Haitz, and M. Ulrich, "A novel camera-to-robot calibration method for vision-based floor measurements," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XI-2-2026, pp. 705–712, 2026.
2. R. K. Lenz and R. Y. Tsai, "Calibrating a cartesian robot with eye-on-hand configuration independent of eye-to-hand relationship," in *Proceedings of the CVPR '88: The Computer Society Conference on Computer Vision and Pattern Recognition*, 1988.
3. R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, 1989.
4. F. C. Park and B. Martin, "Robot sensor calibration: Solving AX = XB on

the Euclidean group," *IEEE Transactions on Robotics and Automation*, vol. 10, 1994.

5. K. Daniilidis, "Hand-eye calibration using dual quaternions," *The International Journal of Robotics Research*, vol. 18, 1999.
6. C. Steger, M. Ulrich, and C. Wiedemann, *Machine Vision Algorithms and Applications*. Wiley-VCH, 2018.
7. M. Ulrich and C. Steger, "Hand-eye calibration of SCARA robots using dual quaternions," *Pattern Recognition and Image Analysis*, vol. 16, no. 1, 2016.
8. M. Ulrich and M. Hillemann, "Generic hand–eye calibration of uncertain robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021.
9. M. Ulrich and M. Hillemann, "Uncertainty-aware hand-eye calibration," *IEEE Transactions on Robotics*, vol. 40, 2024.
10. M. Ulrich, C. Steger, F. Butsch, and M. Liebe, "Vision-guided robot calibration using photogrammetric methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 218, 2024.
11. MVTec Software GmbH, "MVTec HALCON," https://www.mvtec.com/products/halcon, Munich, Germany, 2026, accessed September 18, 2026.
12. Leica Geosystems, "Leica Absolute Tracker AT901 and PCMM specs," 2010.
13. IDS Imaging Development Systems GmbH, "uEye U3-31R0CP-C-HQ Rev. 2.2 camera data sheet," https://en.ids-imaging.com/store/u3-31r0cp-rev-2-2.html, 2026, manufacturer data sheet and product page, Accessed September 22, 2026.
14. Jos. Schneider Optische Werke GmbH, "TOPAZ 2.8/12 C lens data sheet," https://schneiderkreuznach.com/en/industrial-optics/lenses/c-mount-lenses/topaz/f2-8-12mm-c, 2026, manufacturer data sheet, Accessed September 22, 2026.
15. IDS Imaging Development Systems GmbH, "uEye U3-36P0XCP-C-HQ Rev. 1.2 camera data sheet," https://en.ids-imaging.com/store/u3-36p0xcp-rev-1-2.html, 2026, manufacturer data sheet and product page, Accessed September 22, 2026.
16. Jos. Schneider Optische Werke GmbH, "CITRINE 1.8/4.8 C-S lens data sheet," https://schneiderkreuznach.com/en/industrial-optics/lenses/c-mount-lenses/citrine/f1-8-4-8mm-c-stabilized, 2026, manufacturer data sheet, Accessed September 22, 2026.

Single-shot deflectometry using an inverse crossed fringe pattern

Christian Kludt[1] and Jürgen Beyerer[12]

[1] Fraunhofer Institute of Optronics,
System Technologies and
Image Exploitation (IOSB),
Fraunhoferstr. 1, 76131 Karlsruhe
[2] Karlsruhe Institute of Technology,
Kaiserstr. 12, 76131 Karlsruhe

Abstract Phase Measuring Deflectometry (PMD) enables highly sensitive defect detection on specular surfaces but is limited by fringe pattern compression on strongly curved geometries and by the need for multiple phase-shifted exposures. We address both challenges by combining inverse fringe patterns and Fourier-domain single-shot phase retrieval. Using experimentally determined camera-screen correspondences, inverse fringe patterns are computed such that their reflections appear undistorted and spatially periodic in the camera image, without requiring explicit system calibration. Based on these predistorted patterns, a single crossed-fringe image is evaluated using Fourier-domain phase retrieval with single-sideband demodulation (SSD), enabling robust phase extraction from a single exposure. The proposed method eliminates phase shifting, reduces acquisition time, improves robustness against motion and vibration, and extends deflectometry to complex surface geometries and dynamic inspection scenarios.

Keywords Single-shot deflectometry, inverse fringe pattern, Fourier-domain phase retrieval, computational imaging, specular surface inspection, high-speed optical inspection, surface gradient measurement

1 Introduction

PMD is a widely used optical measurement technique for specular surface inspection. In conventional PMD, the two orthogonal gradient components are obtained by displaying and capturing sequences of phase-shifted sinusoidal fringe patterns. Although this approach provides high measurement accuracy, it suffers from two fundamental limitations.

First, surfaces with steep slopes and strong curvatures lead to severe distortion and spatial compression of the reflected fringe patterns in the camera image. The observed fringes become locally dense and irregular. If the fringe frequency exceeds the spatial resolution of the imaging sensor, reliable phase extraction is no longer possible. This limitation restricts the applicability of conventional PMD for objects with complex geometries that are frequently encountered in industrial inspection.

Second, phase retrieval in PMD typically relies on phase-shifting techniques requiring multiple sequential exposures. As a result, the measurement process is inherently time-consuming and sensitive to object motion, vibrations, and environmental disturbances. These limitations make conventional PMD unsuitable for high-speed measurements and dynamic inspection scenarios.

In this work, we propose a single-shot deflectometry method based on inverse fringe patterns. The key idea is to predistort the displayed fringe pattern such that, after reflection at the nominal object geometry, the fringe pattern observed by the camera becomes straight and spatially periodic. In contrast to conventional PMD, geometric distortions caused by the nominal object shape are compensated before image acquisition.

This has two important consequences. First, fringe compression is avoided, resulting in improved spatial sampling across the entire field of view. Second, the observed fringe pattern closely resembles an ideal carrier signal, providing optimal conditions for Fourier-domain phase retrieval. Consequently, phase information can be extracted from a single camera frame using Fourier-domain phase retrieval and SSD.

2 Related Work

Conventional phase-measuring deflectometry relies on phase-shifting techniques requiring multiple exposures for phase retrieval. To overcome this limitation, several single-shot approaches have been proposed. A fundamental contribution is the Fourier-transform method introduced by Takeda et al. [1], which enables phase retrieval from a single fringe image by isolating a carrier component in the frequency domain. Combined with SSD, the method allows efficient extraction of phase information without phase shifting. Liu et al. [2] demonstrated its applicability to deflectometry using crossed fringe patterns.

A common challenge of Fourier-based methods is their dependence on a well-defined carrier frequency and sufficient spectral separation. Surface-induced distortions alter the local fringe frequency and orientation, causing spectral broadening and overlap of the sidebands with the zero-order component. As a consequence, phase retrieval becomes increasingly unreliable for strongly curved or freeform specular surfaces. To address this problem, several single-shot deflectometry methods have been developed that operate directly on crossed fringe patterns. Li et al. [3] combined crossed fringes with Fourier-domain demodulation after background and modulation normalization. Nguyen et al. [4] proposed a one-shot method based on first- and second-order derivatives of the fringe image under the assumption of slowly varying background illumination and modulation. For defect detection applications, Pösch et al. [5] analyzed local fringe orientations using a short-time discrete Fourier transform and subsequent skeletonization. While these approaches eliminate phase shifting, their performance remains affected by geometric fringe distortions caused by the object shape.

A complementary line of research aims at compensating such distortions through pattern adaptation. If information about the nominal object geometry is available, the displayed pattern can be modified such that a desired fringe structure is observed in the camera image. Massig et al. [6] used predistorted crossed fringes to compensate the reflection geometry of the human eye, resulting in approximately regular fringes in the captured image. However, their approach was demonstrated only for simple near-spherical geometries. For more general objects, Li et al. [7] generated inverse fringe patterns from experimentally determined camera-screen correspondences obtained by phase-

measuring deflectometry. Werling and Beyerer [8] followed a similar inversion strategy for specular surface inspection, while Strohmeier et al. [9] demonstrated Fourier-based evaluation of inverse patterns for solely one-dimensional fringe structures.

The method proposed in this paper builds upon these two research directions. Similar to previous work on inverse fringe patterns, the displayed pattern is adapted to the nominal object geometry using experimentally determined camera-screen correspondences. In contrast to existing approaches, the resulting reflected fringe image is specifically designed to satisfy the requirements of Fourier-domain phase retrieval of crossed fringes. The inverse pattern compensates geometry-induced fringe distortion before image acquisition, thereby preventing fringe compression and improving spectral separation in the Fourier domain. This enables robust single-shot phase retrieval using SSD even for strongly curved and freeform specular surfaces.

The main contributions of this work are:

- A fast and robust algorithm for generating inverse fringe patterns from experimentally determined camera-screen correspondences.
- A calibration-free black-box implementation that does not require explicit knowledge of intrinsic or extrinsic system parameters.
- A single-shot deflectometry method based on inverse crossed fringe patterns and Fourier-domain phase retrieval.

3 Proposed Method

When a pattern displayed on a screen is observed via a specular surface, its appearance in the camera image is generally distorted by the object's geometry. This effect can be exploited in reverse. Assuming that a deflectometric reference measurement of the test object is available, inverse fringe patterns can be generated such that, after reflection at the object surface, a predefined fringe pattern appears in the camera image. The displayed pattern therefore contains a compensating geometric distortion which cancels the distortion introduced by the object's shape.

For this purpose we use the phase shifting method [10]. It is extensively used in deflectometry [11] for its robustness against the transmission impairments attenuation, blurring and noise. In the following, we introduce the subscripts $_\mathrm{c}$ and $_\mathrm{s}$ to indicate camera and screen quantities, respectively. To this end, each screen pixel position $\mathbf{x}_\mathrm{s}$ is encoded via a temporal sequence of discrete gray values, the so-called phase shifting sequence. Recording and decoding it yields the deflectometric registration $\boldsymbol{\xi}$, a function

$$\boldsymbol{\xi} : \mathcal{C} \mapsto \mathcal{S} \cup \varnothing, \quad \mathcal{C} \subseteq \mathbb{Z}^2, \quad \mathcal{S} \subseteq \mathbb{R}^2 \tag{1}$$

from the two-dimensional set of all camera pixels $\mathcal{C}$ to the two-dimensional set of screen positions $\mathcal{S}$ (with the empty set representing no interception), yielding estimated screen positions $\hat{\mathbf{x}}_\mathrm{s}$ for every camera pixel $\mathbf{x}_\mathrm{c}$:

$$\boldsymbol{\xi}(\mathbf{x}_\mathrm{c}) = \hat{\mathbf{x}}_\mathrm{s}, \quad \text{with} \quad \mathbf{x}_\mathrm{c} = \begin{pmatrix} x_\mathrm{c} \\ y_\mathrm{c} \end{pmatrix}, \quad \hat{\mathbf{x}}_\mathrm{s} = \begin{pmatrix} \hat{x}_\mathrm{s} \\ \hat{y}_\mathrm{s} \end{pmatrix}. \tag{2}$$

In the same way an image I_α can be modeled as a function

$$I_\alpha : \Omega_\alpha \mapsto \mathcal{I}, \quad \mathcal{I} \subseteq \mathbb{R} \tag{3}$$

from the support points Ω_α to image values $\mathcal{I}$, assigning an image value i_α to every position $\mathbf{x}$, with $\alpha \in \{\mathrm{c}, \mathrm{s}\}$:

$$I_\alpha(\mathbf{x}) = i_\alpha. \tag{4}$$

During a deflectometric measurement an image I_s is displayed on the screen, reflected by the test object and captured with a camera as the image I_c. Then $\boldsymbol{\xi}$ is the geometric transformation of the scene which maps I_s to I_c:

$$I_\mathrm{c}(\mathbf{x}_\mathrm{c}) = I_\mathrm{s}(\boldsymbol{\xi}(\mathbf{x}_\mathrm{c})). \tag{5}$$

In order to generate an image I'_s which compensates $\boldsymbol{\xi}$, we need its inverse function $\boldsymbol{\xi}^{-1} : \mathcal{S} \mapsto \mathcal{C}$. However, this presents an ill-posed problem for the following reasons:

- $\boldsymbol{\xi}^{-1}$ is not well defined because there may be screen positions $\mathbf{x}_\mathrm{s}$ where no camera sight ray is looking at, so that $\boldsymbol{\xi}$ is not surjective.
- $\boldsymbol{\xi}^{-1}$ is not uniquely defined because there might be more than one camera sight ray looking at the same screen positions $\hat{\mathbf{x}}_\mathrm{s}$, so that $\boldsymbol{\xi}$ is not injective. In fact, the inverse is a multivalued function $\boldsymbol{\xi}^{-1} : \mathcal{S} \rightrightarrows \mathcal{C} \cup \varnothing$.
- $\boldsymbol{\xi}^{-1}$ is not continuous: although screen positions $\hat{\mathbf{x}}_\mathrm{s}$ are decoded with continuous values, the two-dimensional set of discrete screen pixels $\tilde{\mathcal{S}}$ can only be activated at discrete pixels $\mathbf{x}_\mathrm{s} \in \tilde{\mathcal{S}} \subseteq \mathbb{Z}^2$.

Because the registration $\boldsymbol{\xi}$ is neither surjective nor injective, it follows that it is not bijective, hence it is not invertible.

As a regularization constraint, we assume a certain degree of smoothness. While this assumption may be violated for objects containing steep slopes or discontinuities, the inverse mapping $\boldsymbol{\xi}^{-1}$ is assumed to be piecewise continuous within local neighborhoods. As a local support region we use a circular neighborhood with radius r chosen according to the projected circle of confusion of a camera sight ray on the screen. With this, we assign an interpolated gray value i'_s to each screen pixel $\mathbf{x}_\mathrm{s}$ as follows:

First, the Euclidean distance between an integer screen pixel position $\mathbf{x}_\mathrm{s}$ and a decoded screen position $\hat{\mathbf{x}}_\mathrm{s} = \boldsymbol{\xi}(\mathbf{x}_\mathrm{c})$ is denoted by

$$d(\mathbf{x}_\mathrm{s}, \hat{\mathbf{x}}_\mathrm{s}) = \|\mathbf{x}_\mathrm{s} - \hat{\mathbf{x}}_\mathrm{s}\|. \tag{6}$$

Second, the set of all camera pixels $\mathbf{x}_\mathrm{c}$ where d evaluates to zero is denoted by

$$\mathcal{C}^0(\mathbf{x}_\mathrm{s}) = \{\mathbf{x}_\mathrm{c} \in \mathcal{C} \mid d(\mathbf{x}_\mathrm{s}, \hat{\mathbf{x}}_\mathrm{s}) = 0\}. \tag{7}$$

Third, the set of all camera pixels $\mathbf{x}_\mathrm{c}$ where d is within the radius r is denoted by

$$\mathcal{C}^r(\mathbf{x}_\mathrm{s}) = \{\mathbf{x}_\mathrm{c} \in \mathcal{C} \mid 0 < d(\mathbf{x}_\mathrm{s}, \hat{\mathbf{x}}_\mathrm{s}) \leq r\}. \tag{8}$$

Fourth, we apply inverse distance weighting [12] using the modified [13] Shepard's method with the weights

$$w(\mathbf{x}_\mathrm{c}) = \begin{cases} d(\mathbf{x}_\mathrm{s}, \hat{\mathbf{x}}_\mathrm{s})^{-2}, & \text{if} \quad d \leq r, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Given an image I_c with image values i_c which we want to capture with the camera, we compute the pixel values i'_s of the converted image by setting the value of the screen pixel $\mathbf{x}_\mathrm{s}$ to the mean value of all image values i_c where the distance d evaluates to zero; the inverse distance weighted mean of the pixel values where the distance is larger than zero but smaller than or equal to the radius r; and zero if no distance value is within the radius r:

$$i'_\mathrm{s} = \begin{cases} \frac{\sum_{\mathbf{x}_\mathrm{c} \in \mathcal{C}^0} I_\mathrm{c}(\mathbf{x}_\mathrm{c})}{|\mathcal{C}^0|}, & \text{if} \quad |\mathcal{C}^0| > 0 \\ \frac{\sum_{\mathbf{x}_\mathrm{c} \in \mathcal{C}^r} w(\mathbf{x}_\mathrm{c}) I_\mathrm{c}(\mathbf{x}_\mathrm{c})}{\sum_{\mathbf{x}_\mathrm{c} \in \mathcal{C}^r} w(\mathbf{x}_\mathrm{c})} & \text{if} \quad |\mathcal{C}^r| > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

When using a *kd-tree* as a fast spatial search structure, this becomes an efficient interpolation method with computational complexity $\mathcal{O}(|\tilde{\mathcal{S}}| \log(|\mathcal{C}|))$ for queries.

A significant advantage of this procedure is that it does not require explicit calibration of either camera or screen. The complete system is treated as a black box, and all required information is obtained experimentally. As a result, system setup is simplified and automated deployment becomes possible.

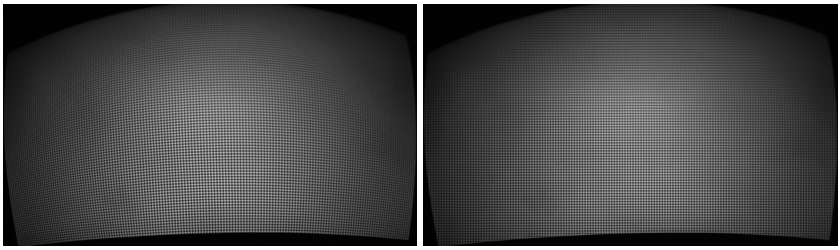
4 Experimental Setup

The experimental setup consists of a flat-panel display for fringe pattern projection, a monochrome camera for image acquisition, and a workstation for inverse fringe pattern generation and phase evaluation. A convex mirror was used as a representative specular test object exhibiting pronounced curvature and strong geometric distortion of reflected fringe patterns.

First, a conventional phase-measuring deflectometric reference measurement was performed to determine the camera-screen correspondence ξ. The acquired registration was subsequently used to generate inverse fringe patterns according to the method described in Section 3. No explicit geometric calibration of the camera, screen, or object was required, as the inversion procedure operates directly on the experimentally determined correspondence data.

During measurement, the inverse crossed fringe pattern was displayed on the screen and observed through reflection at the specular test object, cf. Fig. 1. The carrier frequencies of the conventional and inverse fringe patterns were chosen such that both reflected patterns exhibit identical spatial frequencies in the camera image.

Finally, the recorded images were processed using Fourier-based SSD to recover the phase information.

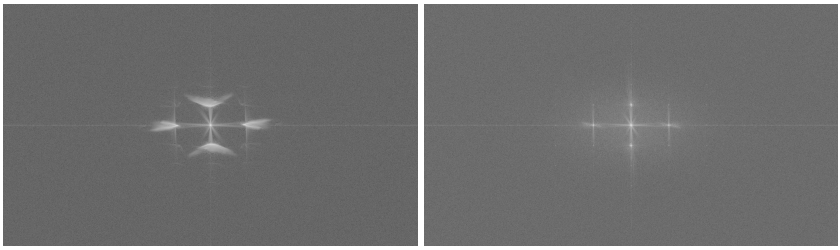


(a) distorted **(b)** inverse

Figure 1: Crossed fringe patterns, both with a spatial frequency $\nu \approx (12 pixel)^{-1}$. The inverse pattern appears undistorted and geometrically regular after reflection.

5 Results and Discussion

The generated inverse fringe patterns successfully compensate the dominant geometric distortions introduced by the nominal object shape. Figure 1 shows that the reflected inverse fringe pattern remains straight and spatially uniform over the entire field of view. In contrast, the conventional fringe pattern exhibits pronounced local compression and distortion caused by the surface curvature of the test object.



(a) distorted **(b)** inverse

Figure 2: Fourier spectra (absolute values, logarithmic scale) of the crossed fringe patterns. The spectrum of the inverted pattern exhibits stronger and more distinct peaks.

The improved spatial regularity is directly reflected in the Fourier domain. Figure 2 compares the spectra of conventionally acquired and inverse fringe patterns. For the conventional measurement, geometric fringe distortions broaden the sidebands and reduce their separation from the low-frequency and DC-components. In contrast, the inverse fringe pattern produces well localized spectral peaks and a significantly narrower spectral distribution.

These spectral properties are particularly advantageous for SSD. The improved separation of the carrier frequencies enables more selective frequency-domain filtering and reduces leakage from neighboring spectral components. Consequently, phase retrieval becomes more robust and less sensitive to geometric distortions of the object under investigation.

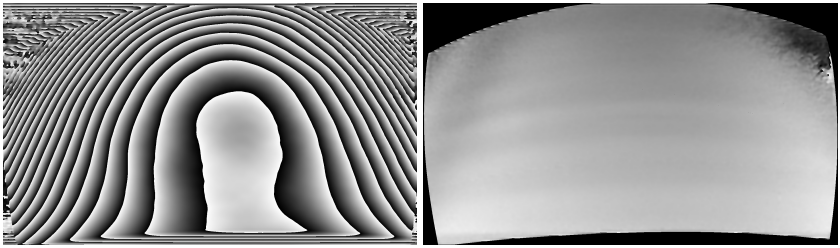
The displayed crossed fringe pattern is characterized by its spatial carrier frequency. After reflection at the specular test object, this carrier is locally modulated by the object's shape and surface slopes, resulting in spatially varying fringe frequencies and distortions.

In contrast, the proposed inverse fringe pattern is pre-distorted such that the modulation introduced by the nominal object geometry is compensated during reflection. Consequently, the fringe pattern observed by the camera exhibits an approximately constant carrier frequency across the field of view.

Figure 3 shows the phase maps recovered from a single camera exposure. After shifting the selected sideband to the frequency origin

for SSD, the phase recovered from the conventional fringe pattern contains the expected phase wraps corresponding to the object-induced modulation, whereas the inverse fringe pattern yields a largely flattened phase distribution because the dominant form contribution has already been compensated during pattern generation.

The results demonstrate that compensating geometric fringe distortions before acquisition establishes favorable conditions for Fourier-domain phase retrieval and enables reliable single-shot deflectometric evaluation.



(a) distorted **(b)** inverse

Figure 3: Fourier-domain phase retrieval with single-sideband demodulation (SSD). The values range from 0 (black) to 2π (white).

6 Conclusion

This paper presented a single-shot deflectometry method that combines inverse crossed fringe patterns with Fourier-domain phase retrieval. Using experimentally determined camera-screen correspondences, inverse crossed fringe patterns are generated such that their reflections appear spatially periodic and geometrically regular in the camera image, without requiring explicit geometric calibration of the measurement system.

The experimental results demonstrate that compensating the nominal object geometry prior to image acquisition substantially reduces fringe compression on strongly curved specular surfaces and improves the spectral separation of the carrier frequencies. As a consequence, the acquired fringe images provide favorable conditions for SSD, enabling

robust phase retrieval from a single exposure. The proposed method therefore combines the geometric advantages of inverse fringe compensation with the speed and simplicity of Fourier-based single-shot evaluation.

References

1. M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry," *Journal of the Optical Society of America*, vol. 72, no. 1, p. 156, Jan. 1982. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=josa-72-1-156
2. Y. Liu, E. Olesch, Z. Yang, and G. Häusler, "Fast and accurate deflectometry with crossed fringes," *Advanced Optical Technologies*, vol. 3, no. 4, pp. 441–445, Aug. 2014. [Online]. Available: https://www.frontiersin.org/articles/10.1515/aot-2014-0032
3. W. Li, T. Bothe, M. K. Kalms, C. Von Kopylow, and W. P. O. Jueptner, "Applications of inverse pattern projection," W. Osten, M. Kujawinska, and K. Creath, Eds., Munich, Germany, May 2003, p. 492. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.508367
4. M. T. Nguyen, Y.-S. Ghim, and H.-G. Rhee, "One-shot deflectometry for high-speed inline inspection of specular quasi-plane surfaces," *Optics and Lasers in Engineering*, vol. 147, p. 106728, Dec. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0143816621001986
5. A. Pösch, T. Vynnyk, and E. Reithmeier, "Using inverse fringe projection to speed up the detection of local and global geometry defects on free-form surfaces," P. J. Bones, M. A. Fiddy, and R. P. Millane, Eds., San Diego, California, USA, Oct. 2012, p. 85000B. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.928700
6. J. H. Massig, E. Lingelbach, and B. Lingelbach, "Videokeratoscope for accurate and detailed measurement of the cornea surface," *Applied Optics*, vol. 44, no. 12, p. 2281, Apr. 2005. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=ao-44-12-2281
7. W. Li, T. Bothe, W. Osten, and M. Kalms, "Object adapted pattern projection—Part I: generation of inverse patterns," *Optics and Lasers in Engineering*, vol. 41, no. 1, pp. 31–50, Jan. 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0143816602001161

8. S. Werling and J. Beyerer, "Automatische Inspektion spiegelnder Oberflächen mittels inverser Muster (Inspection of Specular Surfaces with Inverse Patterns)," *tm - Technisches Messen*, vol. 74, no. 4, pp. 217–223, Apr. 2007. [Online]. Available: https://www.degruyterbrill.com/document/doi/10.1524/teme.2007.74.4.217/html
9. M. Strohmeier, M. Schröder, and C. Faber, "Optical inline inspection detecting 3D defects on complex free-form surfaces in harsh production environments," *tm - Technisches Messen*, vol. 86, no. 6, pp. 335–344, May 2019. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/teme-2019-0013/html
10. M. Servin, J. A. Quiroga, and J. M. Padilla, Eds., *Fringe Pattern Analysis for Optical Metrology: Theory, Algorithms, and Applications*, 1st ed. Wiley, Jul. 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527681075
11. J. Burke, A. Pak, S. Höfer, M. Ziebarth, M. Roschani, and J. Beyerer, "Deflectometry for specular surfaces: an overview," *Advanced Optical Technologies*, vol. 12, p. 1237687, Jul. 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/aot.2023.1237687
12. D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM national conference on -*. Not Known: ACM Press, 1968, pp. 517–524. [Online]. Available: http://portal.acm.org/citation.cfm?doid=800186.810616
13. R. J. Renka, "Multivariate interpolation of large sets of scattered data," *ACM Transactions on Mathematical Software*, vol. 14, no. 2, pp. 139–148, Jun. 1988. [Online]. Available: https://dl.acm.org/doi/10.1145/45054.45055

LineAP: A segment-based average precision metric for visual rail track detection

Annika Bätz and Martin Lauer

Karlsruhe Institute of Technology (KIT),
Institute of Measurement and Control Systems (MRT),
Engler-Bunte-Ring 21, Building 40.32,
76131 Karlsruhe, Germany

Abstract Reliable rail track detection is a prerequisite for safe automated train operation, yet evaluating it requires assessing two distinct aspects of prediction quality: (i) the geometric accuracy of predicted lines, and (ii) the correct instance-level grouping of polylines, e.g. accurate start and end points at railway switches. Existing metrics for polyline detection assess both subtasks simultaneously, losing insight into the error sources of a detection method: partially incomplete but geometrically accurate detections are penalized, while entire rail instances with inaccurate sections are rewarded. We propose LineAP, a segment-based average precision metric that assesses geometric accuracy independently from instance-level correctness by partitioning polyline predictions into fixed-length segments, each evaluated via joint distance and orientation thresholds. A novel track-width-based distance threshold further customizes LineAP to the safety requirements of railway applications, where geometric accuracy remains highly relevant even at large distances from the camera.

Keywords Rail track detection, polyline detection, evaluation metric, average precision, railway perception, lane detection

1 Introduction

Automated train operation requires reliable perception of the train's surroundings. A prerequisite for many downstream tasks – such as ego-path detection, obstacle recognition, scene understanding, and

map-based self-localization – is the precise localization of the railway tracks in camera images.

Railway tracks can be efficiently represented as polylines that mark the left and right rail of each track [1–3] (see Fig. 1). Due to this representation, rail track detection can be formulated as a 2D polyline detection task that is similar to the detection of road lane markings in the automotive domain. As a result, commonly used metrics for rail track detection originate from this domain, such as the metrics from the TuSimple [4] and CULane [5] benchmarks. However, all metrics commonly used for the evaluation of either rail track detection or lane marking detection, share two shortcomings.

First, the metrics treat entire predicted and ground truth instances as the unit of evaluation. As a result, errors such as incomplete or partially correct predictions, and one-to-many or many-to-one correspondences between predictions and ground truth, are not properly accounted for: a prediction is treated entirely as a true positive (TP) or a false positive (FP), regardless of the geometric precision of its individual parts.

Second, existing metrics for 2D polyline detection use fixed distance thresholds in pixel space, and disregard the fact that the same distance error can have different implications depending on the distance of the line object from the camera. As precise rail track detection in the far distance is particularly important for safe train operation, this can lead to misleading evaluation results. A natural alternative would be to evaluate polyline detections in bird's-eye view (BEV) using calibrated real-world coordinates; however, camera calibration is not available for most of the existing rail track detection datasets, such as the ones used in this work, RAIL-BENCH [1] and RailSem19 [2]. Moreover, a BEV projection additionally relies on a flat-world assumption, which not always holds.

To address these issues, we propose LineAP, a segment-based average precision metric that evaluates polyline detections at a finer geometric granularity than existing metrics. Although designed for rail track detection, it can be applied to any polyline prediction task such as road lane marking detection. With the addition of a variable distance threshold that approximates a fixed real-world distance based on the track width, LineAP can be further adapted to the specific requirements of rail track detection.

We demonstrate the benefits of LineAP in a case study on three



Figure 1: Example of rail track annotations in the RAIL-BENCH Rail dataset. Each colored line represents a separate rail instance. Rails of the same track are visualized in the same color. Best viewed in color.

network architectures finetuned on the novel RAIL-BENCH Rail dataset [1], where LineAP yields complementary rankings to the established ChamferAP metric. Together they provide a richer diagnostic than either metric alone.

2 Metrics for Polyline Detection Evaluation

We give a brief, non-exhaustive overview of existing metrics for polyline detection evaluation commonly used in the automotive and railway domains, focusing on those most relevant to our work.

In the TuSimple protocol, ground truth and predicted lines are sampled at uniformly distributed vertical image positions; a ground truth line counts as detected if the horizontal distance to a predicted line stays below a pixel threshold for 85% of these positions. TuSimple Accuracy is the fraction of correctly detected ground truth lines.

For CULane, ground truth and predicted lines are rendered as binary masks (drawn with a fixed line width) and matched via the Hungarian algorithm using pairwise IoU with a fixed threshold of 0.5; the result-

ing TP, FP, and false negative (FN) counts yield the CULane F1 score.

The Chamfer Distance (CD) – the symmetric mean nearest-neighbor distance between uniformly sampled points on two polylines – is a common choice for evaluating polyline predictions within an Average Precision (AP) framework in HD Map Construction [6], which we refer to as ChamferAP; unlike the previous metrics, it is designed for bird's-eye-view (real-world) coordinates. Matching is performed greedily, following the confidence ranking of the predictions and a distance threshold for the CD.

Note that all three metrics operate on the instance level, treating entire predicted and ground truth lines as the unit of evaluation – making it difficult to reward geometrically accurate but incomplete predictions, or to handle one-to-many or many-to-one correspondence errors. In addition, all three use a fixed pixel distance threshold that does not correspond to a consistent real-world distance (ChamferAP included, when applied to 2D image-plane predictions).

3 Track-width-based Distance Threshold

In order to define a variable distance threshold in image space that approximates a fixed real-world distance, we can exploit the fact that the track width is a fixed real-world constant that can be measured – at least approximately – in the image.

We define the track width t_w as the distance between the outside edges of the rail heads as this corresponds to the annotations in the RAIL-BENCH Rail dataset. This width combines the track gauge t_g and the width of the two rail heads t_r:

$$t_w = t_g + 2 \cdot t_r \tag{1}$$

The track gauge t_g is a fixed norm, while the width of the rail heads t_r varies depending on the rail type but can be assumed constant within a single image [7].

To measure the track width in an image, we use a simple heuristic that measures the horizontal distance Δw_k between corresponding rails at uniformly sampled image rows v_k (see Fig. 2). The measurements are then averaged for each image row ($\overline{\Delta w_k}$) and linear regression is applied to obtain a linear function $t_w(v)$ that approximates the track

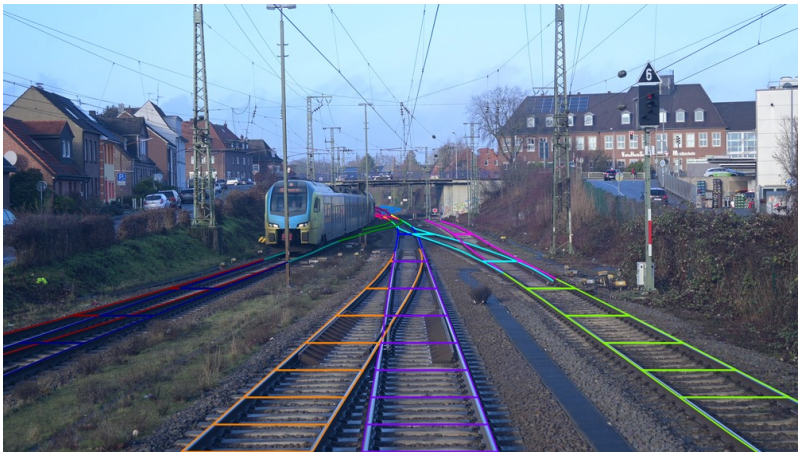


Figure 2: Track width measurements in an image from the RAIL-BENCH Rail dataset. At ten uniformly distributed image rows, the track width is measured as the horizontal distance between the outside edges of the rail heads. Note that the railway sleepers, which are perpendicular to the rails and therefore representative for the track width, are nearly parallel to the horizontal measurement lines. This supports the validity of the horizontal measurement heuristic.

width in image space as a function of the image row (see top right of Fig. 4):

$$t_w(v) = m \cdot v + b \tag{2}$$

Averaged over all images in the RAIL-BENCH Rail dataset, this per-image linear fit achieves a root-mean-square error of 7.6 px against a mean track width of 414.6 px, i.e. a relative error of about 1.8%, supporting its validity.

A variable distance threshold $\Delta d(v)$ can then be defined as a fraction α of the track width in image space:

$$\Delta d(v) = \max(\alpha \cdot t_w(v), \Delta d_{\min}) \tag{3}$$

A lower limit $\Delta d_{\min}$ is applied to avoid overly strict thresholds in the far distance. With the exception of the lower limit, this distance threshold is proportional to the track width and therefore approximates a fixed real-world distance in image space.



Figure 3: Distance threshold in image space: each ground truth polyline is extended by the corresponding distance threshold to both sides. Left: fixed distance threshold of 38 px. Right: variable distance threshold $\Delta d(v)$ of $\alpha = 5\%$ of the track width with a lower limit d_{min} of 5 px.

Figure 3 displays the difference between a fixed pixel threshold and our proposed track-width-based distance threshold $\Delta d(v)$. Note that the distance thresholds are similarly strict in the foreground, but in the background only the track-width-based distance threshold keeps the threshold boundaries close to the ground truth polylines and prevents overlap with neighboring tracks.

Our proposed variable distance threshold $\Delta d(v)$ can be used in any metric that requires a distance threshold in image space. In our experiments, we implement it both in our novel LineAP metric and in our implementation of ChamferAP.

4 LineAP Metric

LineAP builds on the established Average Precision (AP) framework [8]. Its computation involves three consecutive steps: *(A) instance construction*, *(B) instance matching*, and *(C) AP computation*. The latter step is standard and not specific to LineAP, so we focus on the first two steps that define the core of our metric. Figure 4 shows an illustrative toy example.

4.1 Instance Construction

Both ground truth and predicted polylines are partitioned into short segments of a fixed pixel length (the last segment of each line may be

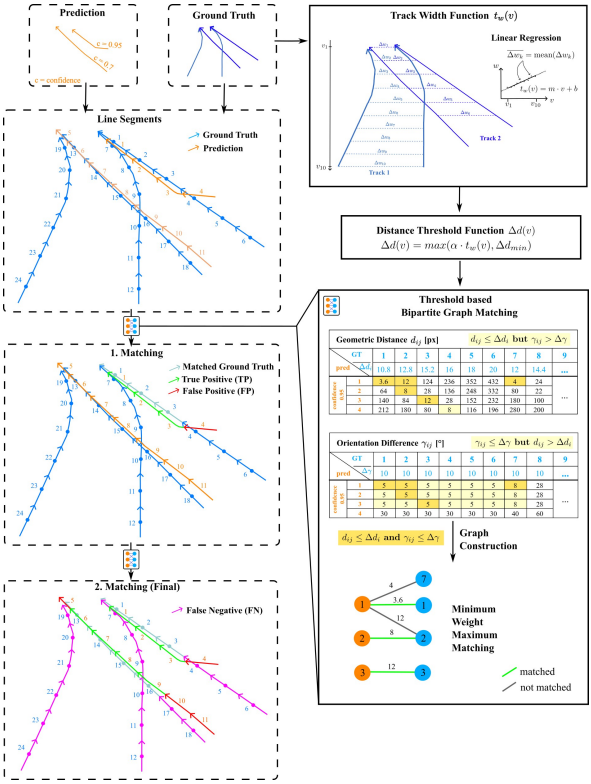


Figure 4: Instance construction and matching in LineAP. Top left: Ground truth (blue) and predicted (orange) polylines are partitioned into short oriented segments. Top right: Construction of the track-width-based distance threshold from the ground truth instances. Lower right: A bipartite graph is constructed based on the variable distance $\Delta d(v)$ and orientation $\Delta\gamma$ thresholds. Minimum-weight maximum matching yields the optimal one-to-one assignment for that confidence level, which is repeated for the next confidence level until all predictions are processed. Best viewed in color.

shorter). Each predicted segment inherits the confidence score of its parent polyline. The resulting ground truth and predicted segments are treated as independent instances that undergo the AP matching process.

4.2 Instance Matching

Matching follows the standard AP protocol: each ground-truth and predicted segment can be matched at most once, and predictions are processed in descending confidence order. For each group of predictions with the same confidence score, we first collect all *valid pairs* of unmatched ground-truth and predicted segments. A pair is valid when two criteria are simultaneously satisfied: (1) the Euclidean distance from the center point of the ground-truth segment to the nearest point on the predicted segment is below the track-width-based distance threshold $\Delta d(v)$, where v is the v-coordinate of the ground truth center point; and (2) the angular difference between the two segment orientations is below another threshold $\Delta\gamma$. This second threshold is fixed across all instances.

Valid pairs are added to a bipartite graph, with nodes representing segments and edges identifying all valid pairs. Minimum-weight maximum matching yields the optimal one-to-one assignment. The process repeats for the next lower confidence group until all predictions are processed. The resulting precision-recall curve is integrated to obtain the AP score (*C: AP computation*).

5 Experiments

5.1 Model Training

We use three different network architectures – PINet [9], YOLinO [10], and CLRNet [11] – to illustrate a range of model capabilities and possible erroneous predictions. All three models are first pretrained on the RailSem19 dataset [2], and then finetuned and evaluated on the novel RAIL-BENCH Rail dataset. We use the predefined train/val/test split of the RAIL-BENCH Rail challenge.

PINet is pretrained and finetuned with a learning rate of 1×10^{-3} for 55 and 35 epochs, respectively. YOLinO is pretrained with a learning rate of 1×10^{-4} for 30 epochs, and then finetuned for 15 epochs with a smaller learning rate of 1×10^{-5}. The raw output of YOLinO, a set of short line segments, is post-processed into the final rail line predictions via greedy graph-based merging of nearby segments. For CLRNet we use a ResNet-18 backbone [12], loss weights of $w_{cls} = 2$, $w_{xytl} = 0.2$,

$w_{LIoU} = 2$, and $w_{seg} = 0.5$, and a non-maximum-suppression threshold of 5. The model is pretrained for 50 epochs with a learning rate of 2×10^{-3}, finetuned on RAIL-BENCH Rail for 40 epochs with a learning rate of 2×10^{-4}.

5.2 Evaluation Settings

In our experiments, we evaluate all models with both LineAP and ChamferAP on the RAIL-BENCH test set. For ChamferAP we keep the general evaluation pipeline, but replace the constant distance threshold with our new track-width-based threshold. For that, we assign to each ground truth line an individual distance threshold using the mean v-coordinate of the ground truth line to compute the track-width-based distance threshold $\Delta d(v)$. The minimal distance threshold is set to 10 px. We use 50 sample points along each predicted and ground truth line to compute the CD.

For LineAP we choose a segment length of 2% of the image width, which ranges between 76 and 109 px, to account for different image resolutions. The orientation threshold is set to 10° and the minimal distance threshold to 5 px. Following standard evaluation protocols, we limit the number of predictions to the 100 with the highest confidence scores for both LineAP and ChamferAP.

6 Results

Table 1 reports the quantitative results. CLRNet achieves the overall highest mAP_{avg}, followed by PINet and YOLinO. Most notably, YOLinO achieves the highest LineAP but the lowest ChamferAP score. As explained in Section 5.1, YOLinO's rail instances are assembled from short line segments via a separate post-processing step, which more often yields incomplete or interrupted lines than PINet or CLRNet – lowering its ChamferAP score. LineAP, however, credits partially correct predictions whenever the predicted segments are sufficiently close to the ground truth and within the orientation threshold, illustrating exactly the diagnostic insight lost when evaluating polyline detections only at the instance level.

Figure 5 shows a qualitative comparison across all three models.

For PINet, only LineAP penalizes the incorrect background predictions. For YOLinO, several incomplete predictions correspond to a single ground truth line and are rewarded only by LineAP. For CLRNet, conversely, two predicted lines correspond to four ground truth lines. ChamferAP still matches both lines to the blue ground truth track in the easy setting, but in the hard setting the CD of the right rail exceeds the threshold and the line is marked incorrect – whereas LineAP fully credits the geometrically correct sections of both rails while penalizing the large gap on the left rail.

Table 1: LineAP and ChamferAP results on the RAIL-BENCH test set. Distance thresholds @α are given in percentage with respect to the track width. mAP is the mean AP across all thresholds for each individual metric and mAP_{avg} is the average between mAP of LineAP and ChamferAP. **Best results** are shown in bold, second-best results are underlined.

	LineAP						ChamferAP						mAP_{avg}
Model	@3	@5	@7	@10	@20	mAP	@5	@10	@20	@30	@50	mAP	
PINet	35.5	45.3	51.0	55.9	63.3	50.2	14.3	27.4	36.8	**40.3**	**43.6**	32.5	41.3
YOLinO	**51.9**	**56.1**	**59.2**	**62.5**	**66.5**	**59.2**	8.8	16.6	22.6	26.7	32.8	21.5	40.4
CLRNet	49.4	53.1	55.7	58.5	62.5	55.9	**22.9**	**32.5**	**37.7**	39.7	41.6	**34.9**	**45.4**

7 Conclusions

Our experiments show that LineAP evaluates polyline predictions at a finer geometric granularity than existing metrics. Combined with an instance-based metric like ChamferAP, we can achieve richer diagnostics of model performance than each metric alone. While LineAP highlights geometric accuracy, an instance-based metric is essential for understanding a model's instance construction capabilities.

Our novel track-width-based distance threshold further allows to customize distance-based metrics to the specific requirements of rail track detection, where precise localization of distant lines is particularly important for safe train operation.

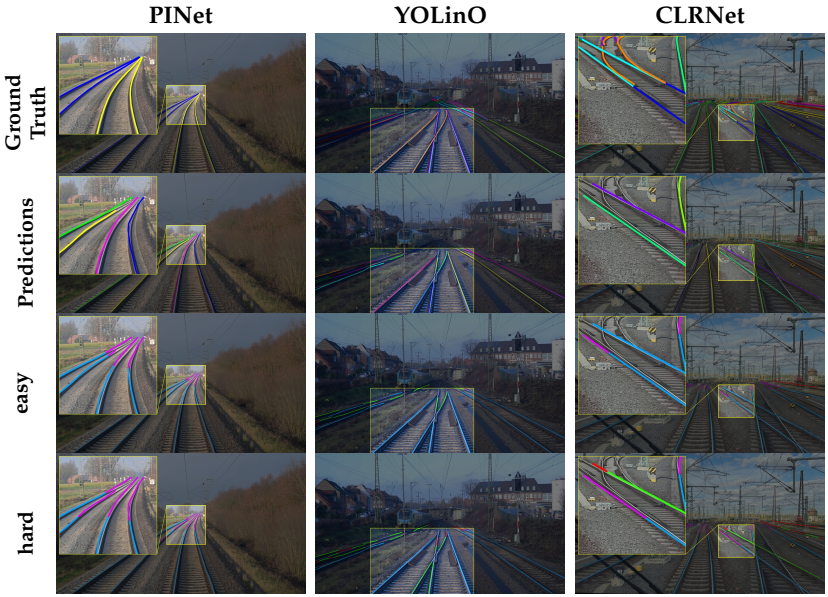


Figure 5: Ground truth (1. row) and predicted rails (2.-4. row) from PINet, YOLinO, and CLRNet, evaluated with LineAP and ChamferAP on RAIL-BENCH Rail in two difficulty levels: easy (LineAP@10 vs. ChamferAP@30) and hard (LineAP@5 vs. ChamferAP@10). Blue: both metrics agree that the prediction is correct. Red: both metrics agree that the prediction is incorrect. Green: identified as correct by LineAP only. Pink: identified as correct by ChamferAP only. Particular interesting cases are highlighted with a yellow box. Best viewed in colors.

Acknowledgements

This work was supported by the Federal Ministry of Transport of Germany within the mFUND program under the grant 01F1178A.

References

1. A. Bätz, P. Klasek, S.-Y. Ham, P. Neumaier, M. Köppel, and M. Lauer, "Railway artificial intelligence learning benchmark (RAIL-BENCH): A benchmark suite for perception in the railway domain," 2026. [Online]. Available: https://arxiv.org/abs/2604.22507

2. O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, and C. Beleznai, "RailSem19: A dataset for semantic rail scene understanding," in *2019 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. Workshops*, 2019, pp. 1221–1229.
3. T. Hiebert, F. Hofstetter, C. Thomas, S. Mushtaq, E. Kaan, and B. Parameswaran, "Labels4Rails: A railway image annotation tool and associated reference dataset," *Data*, vol. 10, no. 12, 2025.
4. S. Yoo, H. Lee, H. Myeong, S. Yun, H. Park, J. Cho, and D. H. Kim, "End-to-end lane marker detection via row-wise classification," 2020, arXiv:2005.08630.
5. X. Pan, X. Zhan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," 2026, arXiv:1712.06080.
6. Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An online HD map construction and evaluation framework," 2022, arXiv:2107.06307.
7. J. Holzfeind, J. Liu, and F. Pospischil, Eds., *Handbuch Eisenbahninfrastruktur*. Springer Berlin Heidelberg, 2025. [Online]. Available: https://link.springer.com/book/10.1007/978-3-662-68960-8
8. M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. of Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
9. Y. Ko, Y. Lee, S. Azam, F. Munir, M. Jeon, and W. Pedrycz, "Key points estimation and point instance segmentation approach for lane detection," *IEEE Trans. on Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8949–8958, 2022.
10. A. Meyer, P. Skudlik, J.-H. Pauls, and C. Stiller, "YOLinO: Generic single shot polyline detection in real time," in *Proc. of the IEEE/CVF Int. Conf. on Comput. Vis. Workshops*, 2021, pp. 2916–2925.
11. T. Zheng, Y. Huang, Y. Liu, W. Tang, Z. Yang, D. Cai, and X. He, "CLRNet: Cross Layer Refinement Network for Lane Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022.
12. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

On-camera person localization with omlox integration for safety applications

Jannick Ahlvers[1], Dennis Sprute[1], Gerrit Holzbach[2], and Holger Flatt[1]

[1] Fraunhofer IOSB, Industrial Automation Branch (IOSB-INA),
Campusallee 1, 32657 Lemgo
[2] Fraunhofer IOSB,
Fraunhoferstraße 1, 76131 Karlsruhe

Abstract Modern industrial production is shifting towards flexible, modular environments in which safety concepts bound to a fixed geometry no longer suffice. This paper presents a localization system as a basis for digital safety zones. It combines two complementary location providers: ultra-wideband (UWB) for tagged assets, and smart embedded cameras for person localization. Each camera acts as an autonomous localization node that performs person detection, multi-object tracking, ground-plane projection and georeferencing entirely on-device, so that raw image data never leave the camera and only abstracted positions are published. This supports GDPR-compliant operation through data minimization and privacy by design. Both providers are registered as equivalent location providers in omlox, an open standard for interoperable indoor localization, which yields a single georeferenced view of all entities. Two NVIDIA Jetson-based smart cameras and a UWB infrastructure with nine satellites were evaluated in the SmartFactoryOWL in Lemgo against reference points. Persons are located at a median error of 0.16 m with a stationary scatter of 0.02 m by the cameras, comparable to the UWB provider (med. error 0.16 m, scatter 0.09 m). Both location provider achieve real-time behavior with 6 Hz and 4 Hz for camera and UWB localization, respectively. The results indicate that on-camera and the complementary UWB tag-based localization can serve as the basis for digital safety zones that are configured in software rather than installed physically.

Keywords Person localization, omlox, embedded vision, industrial safety, edge computing, UWB

1 Introduction

Modern industrial production is shifting from static, line-based factories towards flexible, modular environments with frequently changing products, workflows and layouts. Established safety concepts such as light curtains or stationary safety laser scanners are bound to a predefined geometry and require a costly physical re-parameterization after every layout change [1], while mobile safety sensors and tag-based localization only indicate that an obstacle or a tagged entity is nearby, without covering all persons present or distinguishing humans from other objects [2,3]. Together with requirements regarding real-time behavior and compliance with the EU General Data Protection Regulation (GDPR) [4], these limitations motivate the target system of this work: a localization system that (i) provides the basis for digital safety zones that are configured in software rather than installed physically, (ii) distinguishes persons from other objects, (iii) operates in real time, and (iv) processes personal data in a data-protection-compliant manner.

As the purpose is safety-relevant, a single locating technology is not adequate, since each leaves systematic blind spots under its characteristic failure mode. Ultra-wideband (UWB) localization and cameras are heterogeneous in this sense and additionally complementary: UWB is tag-based and needs no line of sight, covering machines and mobile platforms, whereas persons carrying no tag are localized and classified using mature computer-vision algorithms. LiDAR and radar offer comparable tag-free coverage and support person classification, but at higher sensor cost; therefore, cameras are used in this work. Combined, the tagged and untagged parts of the environment are covered, and requirements (i)–(iii) are addressed. However, requirement (iv) conflicts with camera systems that stream video to a central server, which becomes a single aggregation point of image data of identifiable persons. Such deployments also scale in bandwidth with resolution and camera count, introducing a single point of failure. The proposed person localization pipeline is therefore executed on the capturing device, so that the camera becomes an autonomous localization node that emits only abstracted positions: no raw images leave the device, and personal data remain confined to it and to a single frame, following data minimization and privacy by design. Because both modalities already output positions rather than raw measurements, integrating them re-

duces to publishing all positions in a shared coordinate system through a common interface. This abstraction is specified by omlox [5], the first open standard for real-time localization. Camera nodes and the UWB infrastructure are registered as location providers within one omlox deployment zone, so that consuming applications receive positions from both sources without knowledge of the underlying technology.

Accordingly, an on-device pipeline for person detection, multi-object tracking, ground-plane projection, and georeferencing on a smart embedded camera is contributed, and integrated with a UWB infrastructure into a common omlox deployment zone. The resulting system is evaluated on the shopfloor in the SmartFactoryOWL in Lemgo with respect to requirements (i)–(iv).

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes proposed localization system, Section 4 presents the experimental setup, and Section 5 discusses the results before Section 6 concludes.

2 Related Work

In contrast to protective devices that react to an intrusion into a fixed geometry, localization systems provide the positions of persons continuously. Tag-based radio approaches such as UWB offer robust coverage independent of the line of sight, and safety functions can in principle be derived from real-time UWB localization [6], although accuracy degrades under non-line-of-sight propagation and tags rather than persons are localized. Several fusion schemes therefore combine UWB with camera measurements in a joint estimator, for instance through particle or federated Kalman filtering, reporting accuracies from a few centimeters up to about 0.2 m [7–9]. Comparable accuracies are reported in a joint evaluation of a UWB-based and a camera-based locating system in an operational industrial environment, which also confirms the complementary error characteristics, with median errors of 12–13 cm in static conditions, degrading to 19 cm for the radio-based system under human shadowing and to 25 cm for the camera-based system under mobility [10]. These approaches refine the position of an entity observed by both modalities and thus still presuppose a tag on every localized person, or treat the camera as a backup for a primary

tag-based installation rather than as an equivalent source. Tag-free localization is likewise achieved with LiDAR and radar. For LiDAR, person detection is performed by models trained specifically on point-cloud data [11, 12], whereas indoor radar requires the suppression of multipath ghost targets in metal-rich environments before positions become reliable [13]. Both modalities are therefore feasible, but cameras were selected here because of their lower sensor cost and the broader availability of pre-trained detection models.

Regardless of the sensing modality, camera, LiDAR and radar share a centralized architecture in which data must leave the device, either for a tightly coupled estimator or for inference on a server. This introduces bandwidth scaling with the number of devices, adds network latency, and makes the handling of personal data a property of the overall system [14]. As opposed, distributed smart camera networks execute inference on the camera node and exchange only abstracted results [15], and embedded accelerators make deep-learning-based person detection feasible in real time [16]; they are, however, not concerned with publishing positions in a local coordinate system of the shopfloor through an interface that applications can consume without knowledge of the underlying sensor technology. Interoperable localization interfaces such as omlox specify this abstraction, and their use for functional safety is currently being explored [17]. This work brings both lines together: instead of a camera feeding a central estimator, each camera performs detection and tracking on-device and publishes georeferenced positions through omlox, where it acts as a location provider equivalent to the tag-based UWB infrastructure.

3 Localization System: Camera and UWB

The proposed system combines a camera-based localization infrastructure built from smart embedded cameras with on-board GPUs, a complementary UWB-based localization infrastructure, and an omlox hub that unifies both technologies at the localization layer as shown in Fig. 1.

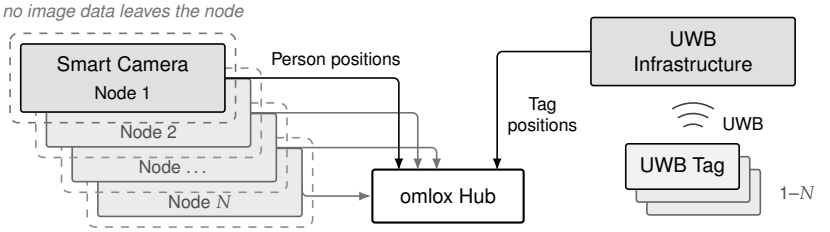


Figure 1: System architecture: Camera nodes and the UWB infrastructure publish positions to a common omlox hub; raw image data remain on the camera nodes.

3.1 Node Concept and Failure Isolation

The central design decision of the presented architecture is that all image processing is performed on the camera itself. No image data are transmitted beyond the device and no central server is involved.

This decentralization has three consequences. First, the network load introduced by a node is independent of the image resolution and frame rate since only position messages are transmitted. The infrastructure therefore scales with the number of nodes without the bandwidth growth. Second, the computational load is distributed by construction: adding a node adds its own processing resources, so no shared inference capacity has to be provisioned. Third, failure of a single node does not affect the remaining nodes. Since each node is functionally self-contained, a node failure reduces the covered area but leaves the localization capability of the overall system intact. This contrasts with centralized architectures, in which the processing server constitutes a single point of failure for all connected cameras.

In addition, decentralization supports data protection requirements. Image data are processed and discarded on the device, and only position objects without appearance information leave the camera. The node thus behaves as a closed component that exposes localization information rather than video, which follows the principles of data minimization and privacy by design [4] and simplifies the deployment of camera-based localization in areas where video surveillance would be difficult to justify.

Integration with the remaining localization infrastructure is achieved through the omlox hub. The camera nodes and the UWB system pub-

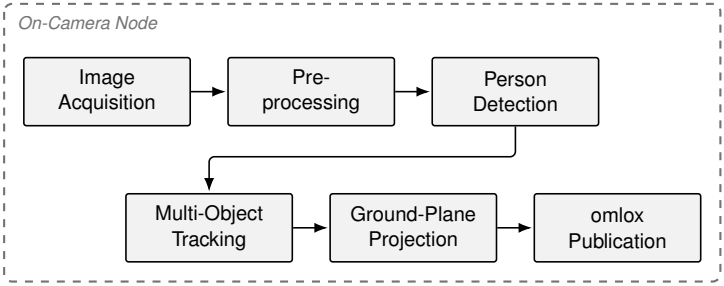


Figure 2: On-device processing pipeline of a camera node.

lish into the same omlox deployment zone and therefore into a common coordinate reference. Consumers of localization data subscribe to the hub and receive position updates without knowledge of the underlying sensing technology, which allows the camera-based localization to be added to or removed from an existing installation without changes on the consumer side.

3.2 Pipeline Stages

The on-device image processing chain consists of six stages that are executed for every frame. Fig. 2 gives an overview of the pipeline; the following subsections describe the individual stages.

Image acquisition and preprocessing. Each camera node captures color images at the native sensor resolution. For person detection, the image is scaled to the detector input resolution, normalized, and transferred to the GPU as a batched tensor.

Person detection. Person detection is performed by a single-stage CNN-based object detector executing on the embedded GPU of the camera, which returns bounding boxes and object classes with confidence scores. Only the person class is retained, so no other object class is published by a node.

Multi-object tracking. Per-frame detections are associated over time using a Kalman filter per track for motion prediction and combines the distance between predicted and observed box positions with an appearance descriptor from a lightweight re-identification network.

Tracking provides temporally consistent identifiers that allow the hub to associate successive position updates with the same person. It also suppresses spurious detections, since a track is only reported after confirmation over several consecutive frames, while tracks are retained for a limited number of frames without matching detections to bridge short detection failures caused by partial occlusion.

Ground-plane projection. The tracked bounding boxes are converted into floor positions under the assumption that persons stand on a planar floor. For each track, the midpoint of the lower bounding box edge is used as the reference point, as it approximates the contact point between person and floor. The reference point is mapped to metric floor coordinates via a planar homography.

Two properties of this approach are relevant for the interpretation of the results. The planarity assumption restricts the method to floor areas without steps or ramps, and the accuracy of the projected position depends on the accuracy of the estimated contact point, which degrades under partial occlusion of the lower body and for shallow viewing angles at large distances from the camera.

omlox publication. The floor positions of all confirmed tracks are published to the omlox hub via its HTTP interface, with the camera node acting as a registered location provider. Since the homography maps directly into the local coordinate system, no additional coordinate transformation is required on the device. Each published message contains the position, the track identifier, and the acquisition timestamp.

3.3 Calibration and Georeferencing

The homography maps image coordinates directly into the local coordinate system, a single calibration step establishes both the image-to-floor mapping and the georeferencing of the node.

Calibration is performed with a set of floor points that are marked in the monitored area and surveyed in the local coordinate system using a laser distance meter. The corresponding image points are determined in a captured frame, and the homography is estimated from these correspondences in a least-squares sense.

The distribution of the calibration points is relevant for the resulting accuracy. Points are placed across the entire monitored area, so that

the estimated homography is not biased towards the well-conditioned image region close to the camera.

4 Experimental Setup

Two camera nodes were deployed in the SmartFactoryOWL in Lemgo, a research production facility with modular workstations and mobile platforms. Node 1 observes an area of approximately $8 \times 15\,\mathrm{m}$ containing a walkway, Node 2 an area of approximately $9 \times 10\,\mathrm{m}$ containing a workstation. The two areas represent the dominant situations of the target application: persons passing through a region and a workstation for human-robot interaction. Node 1 is mounted at a height of 4.7 m with a downward tilt of approximately $18\,^{\circ}$, Node 2 at a height of 6.1 m with a tilt of $44\,^{\circ}$, so that in both cases occlusion of the ground plane is minimized.

Both nodes are Imago VisionCam XM2 devices with an integrated NVIDIA Jetson Orin Nano and acquire images at the native sensor resolution of 2432×2064 pixels with a shutter time of 4.5 ms. Detection uses YOLOv4 [18] with pre-trained COCO weights at an input resolution of 416×416 pixels, executed through TensorFlow 2.x inside the NVIDIA TensorFlow container for integrated GPUs (24.10-tf2-py3-igpu) at FP32 precision; non-maximum suppression uses a score threshold of 0.60 and an IoU threshold of 0.70. Deep SORT [19] is used as multi-object tracker with the `mars-small128` re-identification model, a maximum cosine distance of 0.4 and a maximum IoU distance of 0.7; a track is confirmed after 5 consecutive detections and retained for up to 60 frames without a matching detection. Each homography was estimated from six floor points. At the omlox hub, both nodes and the UWB infrastructure are registered as location providers in a common deployment zone.

The UWB reference system uses nine satellites mounted at 6 m height. Tags are localized at a frequency of 4 Hz.

Two quantities per node were evaluated: processing time and positioning accuracy. The per-stage processing times were logged at a frequency of 1 Hz. The accuracy is evaluated at nine validation points (not used for camera calibration) per node, whose local coordinates were determined using a laser distance meter with a specified accu-

Table 1: Per-stage processing times in milliseconds; total reports the median and 95th percentile (P95) of summed per-frame processing time. (n = 1180).

Pipeline stage	**Median**	**P95**
Image acquisition	10.3	18.9
Preprocessing	5.6	9.6
Person detection	126.7	143.0
Multi-object tracking	13.5	29.5
Ground-plane projection	0.5	0.8
omlox publication	6.8	18.0
Total	**165.7**	**208.2**

racy of ± 2 mm. The accuracy is the Euclidean distance in the ground plane between reported and reference position. During each run, a person wearing a UWB tag visited all nine validation points, being the only person in the scene. The person was oriented towards the camera and stood centrally on the marked reference point with both feet. At each point a stationary stop of at least 30 s was held, with durations varying between points. From each stop the first and the last 4 s were excluded so that movement toward the point and residual position oscillation after stopping do not enter the evaluation, which yields effective windows of 23 s on average.

5 Results

Real-time capability is assessed by throughput and latency. The pipeline stages of Fig. 2 run sequentially per frame, so their sum is the on-device latency from image acquisition to omlox publication. Tab. 1 lists the per-stage times. Person detection accounts for 76 % of the median frame time. The median latency is 165.7 ms, corresponding to 6.0 Hz. The UWB installation publishes at 4.0 Hz.

Localization accuracy is assessed by the deviation from the surveyed reference, aggregated over the nine validation points per node as median, root-mean-squared-error (RMSE) and 95th percentile. In addition, $\bar{\sigma}$ gives the standard deviation of the positions reported at a stationary point. Tab. 2 shows the results for both providers over the same validation intervals, except for one point without camera detections.

Table 2: Localization error for camera and UWB showing the median error, the root-mean-squared-error (RMSE) and the 95th percentile (P95) $\bar{\sigma}$ is the mean per-point standard deviation. All values in meters.

Provider	n	**Median**	**RMSE**	**P95**	$\bar{\sigma}$
Camera	2228	0.160	0.186	0.284	0.023
UWB	1600	0.157	0.203	0.318	0.087

The camera node localizes persons without a tag at a median error of 0.16 m, and thus within the accuracy band of the tag-based and fusion-based systems of Section 2. The 95th percentile of the camera error remains below 0.29 m, so no validation point deviates substantially. In one run one validation point yielded no camera position, as the person was not detected there; the point is among the three most distant ones from the camera and was affected by cast shadow in combination with dark clothing. Since the camera provider relies on visual detection, its coverage depends on illumination, contrast and distance, whereas the UWB provider covered the same point without interruption. Both providers reach comparable absolute accuracy, but with different error structure. The stationary scatter is small for both providers, with $\bar{\sigma} = 0.02$ m for the camera versus 0.09 m for UWB.

6 Conclusions and Future Work

A localization system was developed in which camera nodes and a UWB infrastructure act as equivalent location providers within one omlox deployment zone. Each camera node performs person detection, tracking, ground-plane projection and georeferencing entirely on the device, so that untagged persons are covered alongside the tagged entities of the radio-based infrastructure. Persons were localized without a tag at a median error of 0.16 m, comparable to the UWB provider. This accuracy provides a basis for determining the tolerance margin between digital safety zones and the physical exclusion areas, addressing requirement (i). Using a deep learning-based object detector to distinguish between object classes, requirement (ii) is addressed. With an update rate of 6 Hz per camera node, alongside 4 Hz for the UWB provider, requirement (iii) is satisfied. Since raw images are processed

and discarded on the device and only positions, track identifiers and timestamps are transmitted, requirement (iv) is addressed through data minimization and privacy by design.

Future work concerns the coverage gap caused by missed detections under unfavourable illumination, which motivates a detector adapted to the deployment environment. Beyond this, the evaluation needs to be extended to multiple simultaneously present persons. In addition, the position estimates for a single person from multiple nodes with overlapping fields of view could be combined. Finally, the derivation of safety functions from the published positions is to be examined against the applicable standards.

References

1. International Electrotechnical Commission, "IEC 61496-1:2020, Safety of machinery – Electro-sensitive protective equipment – Part 1: General requirements and tests," International Standard, Edition 4.0, 2020.
2. S. E. Navarro, S. Mühlbacher-Karrer, H. Alagi, H. Zangl, K. Koyama, B. Hein, C. Duriez, and J. R. Smith, "Proximity perception in human-centered robotics: A survey on sensing systems and applications," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1599–1620, 2022.
3. D. Colaiuda, M. Ragnoli, G. Ferri, B. Tambelli, G. Valente, and V. Stornelli, "UWB-based personal protective equipment for hazard detection in working environments," in *Proceedings of SIE 2025*. Cham: Springer Nature Switzerland, 2026, pp. 273–280.
4. European Parliament and Council of the European Union, "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," Official Journal of the European Union, L 119, pp. 1–88, 2016.
5. PROFIBUS Nutzerorganisation e.V. (PNO), "omlox core zone specification, version 2.1.1," PROFIBUS Nutzerorganisation e.V., Karlsruhe, Germany, Technical Specification Order No. 20.122, 2023.
6. W. Wang, Z. Zeng, W. Ding, H. Yu, and H. Rose, "Concept and validation of a large-scale human-machine safety system based on real-time UWB indoor localization," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 201–207.

7. Y. Xu, Z. Chen, M. Zhao, F. Tang, Y. Li, J. Liu, and N. Kato, "UVtrack: Multimodal indoor seamless localization using Ultra-Wideband communication and vision sensors," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 272–281, 2025.
8. F. Liu, J. Zhang, J. Wang, H. Han, and D. Yang, "An UWB/vision fusion scheme for determining pedestrians' indoor location," *Sensors*, vol. 20, no. 4, p. 1139, 2020.
9. P. Peng, C. Yu, Q. Xia, Z. Zheng, K. Zhao, and W. Chen, "An indoor positioning method based on UWB and visual fusion," *Sensors*, vol. 22, no. 4, p. 1394, 2022.
10. S. Valiollahi, I. Rodriguez, W. Zhang, H. Sharma, and P. Mogensen, "Experimental evaluation and modeling of the accuracy of real-time locating systems for industrial use," *IEEE Access*, vol. 12, pp. 75 366–75 383, 2024.
11. D. Basile, D. Sprute, H. Dörksen, and H. Flatt, "GDPR-compliant person recognition in industrial environments using MEMS-LiDAR and hybrid data," arXiv:2602.01764, 2026.
12. D. Basile, D. Sprute, H. Dörksen, and H. Flatt, "Evaluation of 3D-LiDAR based person detection algorithms for edge computing," in *Forum Bildverarbeitung 2024*. KIT Scientific Publishing, 2024, pp. 159–170.
13. J.-K. Park, J.-H. Park, and K.-T. Kim, "Multipath signal mitigation for indoor localization based on MIMO FMCW radar system," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 2618–2629, 2024.
14. T. Bai, H. Zhao, L. Huang, Z. Wang, D. In Kim, and A. Nallanathan, "A decade of video analytics at edge: Training, deployment, orchestration, and platforms," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 2127–2162, 2026.
15. B. Rinner and W. Wolf, "An introduction to distributed smart cameras," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1565–1575, 2008.
16. W. Rahmaniar and A. Hernawan, "Real-time human detection using deep learning on embedded platforms: A review," *Journal of Robotics and Control (JRC)*, vol. 2, no. 6, pp. 462–468, 2021.
17. PROFIBUS Nutzerorganisation e.V., "omlox safety - development in functional safety," https://omlox.com/safety-wg, accessed: 2026-09-11.
18. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
19. N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

Autofocus-based surface reconstruction in digital holography

Harish Sasikumar, Balasubrahmaniyam Mukundakumar, and Shimpei Matsuura

Mitutoyo Research Center Europe B.V.,
De Rijn 18, 5684 PJ Best, The Netherlands

Abstract Digital holography (DH) enables quantitative three-dimensional surface measurement through optical phase reconstruction, but its measurable height range is ultimately constrained by phase wrapping. Autofocus (AF)-based reconstruction has been proposed as an alternative approach in which surface height is estimated from the best-focus position of numerically propagated phase images. In this work, we investigate the factors governing the performance of AF-based reconstruction, with particular emphasis on synthetic wavelength, surface tilt, and surface roughness. Experimental results show that the shape of the AF response and the accuracy of focus localization depend on synthetic wavelength, surface tilt, and surface roughness. A heuristic phase-decorrelation model is introduced to interpret the observed trends and provide guidance for parameter selection. Using the resulting reconstruction framework and parameter choices, height variations exceeding 600 μm are reconstructed with a synthetic wavelength of approximately 14 μm, demonstrating measurement ranges well beyond the wrapping-limited range of conventional phase-based methods.

Keywords Digital holography, autofocus, shape-from-focus, computational imaging, surface metrology, synthetic wavelength

1 Introduction

Three-dimensional surface metrology is essential in applications ranging from precision manufacturing and semiconductor inspection to

biomedical imaging [1–3]. Among optical techniques for non-contact, high-resolution surface characterization, digital holography (DH) is particularly attractive because it provides direct access to the phase of the optical field [4]. The reconstructed complex field also enables numerical refocusing and three-dimensional surface reconstruction without mechanical scanning or repeated measurements [5].

Despite these advantages, quantitative height measurement in DH is fundamentally limited by phase wrapping. Because the measured phase is periodic, different surface heights can produce identical phase values, restricting the unambiguous measurement range. Dual-wavelength DH addresses this limitation by combining measurements acquired at two wavelengths to generate a synthetic phase with a substantially larger wrapping interval [6]. Although this extends the measurable range, it does not eliminate phase ambiguity. Moreover, achieving very large synthetic wavelengths requires increasingly small wavelength separations, imposing stringent requirements on wavelength control and stability.

This limitation motivates the investigation of alternative reconstruction strategies. Autofocus (AF)-based surface reconstruction estimates surface height from the propagation distance corresponding to the best focus rather than directly from phase [7]. DH is particularly well suited to this approach because DH reconstructs the complete complex optical field, enabling numerical propagation to arbitrary axial positions [8].

While AF-based reconstruction has been successfully demonstrated, the factors governing AF response shape and focus-localization performance have not been systematically characterized. In this work, we investigate AF-based surface reconstruction in a dual-wavelength DH [9], with emphasis on the factors governing reconstruction reliability and practical applicability. In particular, we examine the influence of synthetic wavelength, surface tilt, and surface roughness on the AF response and interpret the observed behavior using a heuristic phase-decorrelation model. The model identifies the optical depth of focus (DOF) as a relevant axial length scale and indicates that the synthetic wavelength should be selected relative to the DOF and the required axial sensitivity. Finally, the approach is demonstrated on a representative sample, where height variations exceeding 600 μm are reconstructed using a synthetic wavelength of approximately 14 μm in a system with a DOF of approximately 100 μm.

2 Theory

2.1 Dual Wavelength Digital Holography

Digital holography (DH) reconstructs a complex-valued representation of the wavefront from a recorded hologram. At an illumination wavelength λ, the reconstructed field can be written as

$$E_\lambda(x,y) = A_\lambda(x,y)\exp\left[i\phi_\lambda(x,y)\right], \tag{1}$$

where $A_\lambda(x,y) = \sqrt{I(x,y)}$, $I(x,y)$ is the conventional intensity image, and $\phi_\lambda(x,y)$ denotes the phase distribution.

The phase is proportional to the surface height and therefore enables quantitative surface reconstruction. However, because the phase is periodic, different surface heights can produce identical phase values beyond the phase-wrapping interval, limiting the measurable height range. A strategy for increasing the measurable range is to combine phase information acquired at two illumination wavelengths [6]. In dual-wavelength DH, the reconstructed fields are combined to form a synthetic field ,

$$E_s(x,y) = E_{\lambda_1}(x,y) \circ E^{\dagger}_{\lambda_2}(x,y), \tag{2}$$

where $\circ$ denotes element-wise multiplication and $\dagger$ denotes complex conjugation. The corresponding synthetic phase is

$$\phi_s = \phi_{\lambda_1} - \phi_{\lambda_2}, \tag{3}$$

which is associated with an effective synthetic wavelength,

$$\Lambda_s = \frac{\lambda_1\lambda_2}{|\lambda_1 - \lambda_2|}. \tag{4}$$

An equivalent description is obtained using the synthetic wavenumber

$$k_s = k_1 - k_2, \qquad k_i = \frac{2\pi}{\lambda_i}. \tag{5}$$

For closely spaced wavelengths, Λ_s can be much larger than either illumination wavelength, resulting in a substantially increased phase-wrapping interval and measurement range.

2.2 Numerical Propagation of the Synthetic Field

The reconstructed fields are numerically propagated using the angular spectrum method [10]. For a propagation distance z_n, the propagated fields $E_{\lambda_1}(x, y; z_n)$ and $E_{\lambda_2}(x, y; z_n)$ are combined according to Eq. 2 to generate the synthetic field $E_s(x, y; z_n)$ from which the synthetic phase (Eq. 3) is obtained for autofocus evaluation.

2.3 Autofocus-Based Height Reconstruction

Following Seyler et al. [8], surface height is estimated from an autofocus (AF) function derived from the local consistency of synthetic-phase gradients. For each propagation distance z_i, the AF function $s_{z_i}(x, y) = |\langle \nabla\phi_s / |\nabla\phi_s| \rangle|^2$ quantifies the local alignment of neighboring phase gradients and acts as a focus measure, reaching a maximum near the correct reconstruction plane.

For a given pixel (x, y), the values of $s_z(x, y)$ evaluated over all propagation distances form an AF curve. The focus position is estimated from the center of mass of the AF curve,

$$h(x, y) = \frac{\sum_n z_n \, \tilde{s}_{z_n}(x, y)}{\sum_n \tilde{s}_{z_n}(x, y)}. \tag{6}$$

Unlike the Gaussian fitting approach of [8], the center-of-mass estimator is computationally simple and does not require iterative curve fitting or assumptions regarding a particular functional form of the AF peak. A potential limitation is its sensitivity to asymmetric AF responses.

As a further modification, an amplitude-based quality mask is applied to suppress unreliable height estimates. The synthetic-field amplitude serves as an SNR-like measure of local reconstruction quality, and pixels with amplitudes below 30% of the maximum synthetic-field amplitude are excluded from the reconstructed height map.

2.4 Heuristic Model of the AF Response

To provide physical intuition for AF-based reconstruction, a simple heuristic framework is introduced to model the influence of synthetic

wavelength, surface roughness, surface tilt, and axial defocus on the local alignment of synthetic-phase gradients. The framework is intended as a qualitative description of how these factors affect phase-gradient consistency and, consequently, the shape of the AF response. The resulting phase-gradient consistency, here referred to as coherence, is approximated as

$$\rho \propto \exp\left[-\left(4(k_s\sigma\cos\alpha)^2 + \frac{1}{2}\left(\frac{k_s}{k_0}\frac{\Delta z}{\mathrm{DOF}}\right)^2\right)\right]. \tag{7}$$

where k_s is the synthetic wavenumber defined in Eq. 5, σ is the RMS surface roughness, α is the surface tilt, $k_0 = (k_1 + k_2)/2$ is the mean wavenumber corresponding to the illumination wavelengths λ_1 and λ_2, Δz is the axial displacement from the best-focus plane, and DOF is the depth of focus.

The first term in the exponent (Eq. 7) describes coherence reduction associated with surface roughness and tilt, whereas the second term describes coherence reduction associated with axial defocus.

The heuristic model suggests several trends that are directly relevant to AF-based reconstruction:

- Increasing surface roughness reduces the consistency of local phase gradients, resulting in broader AF responses and reduced focus-localization accuracy.
- Increasing surface tilt reduces the roughness-induced coherence reduction term and therefore sharpens the AF response. For the relatively small tilt angles considered here, the model predicts only a modest effect. In practice, the influence of tilt can be amplified in low-NA systems due to their increased angular selectivity.
- Increasing the synthetic wavenumber k_s, or equivalently reducing the synthetic wavelength Λ_s (Eq. 4), increases the sensitivity of the AF response to defocus. This dependence is governed by the normalized defocus term $(k_s/k_0)(\Delta z/\mathrm{DOF})$, indicating that the choice of synthetic wavelength should be considered relative to the system depth of focus rather than as an independent parameter. Shorter synthetic wavelengths therefore produce narrower AF peaks and improved focus localization.

In the following section, the influence of synthetic wavelength, surface tilt, and surface roughness on the AF response is investigated experimentally and compared with the trends predicted by this heuristic framework.

3 Experiments, Results and Discussion

The experiments were performed using the dual-wavelength digital holography system described by Matsuura et al. [9]. The effect of synthetic wavelength was studied using multiple wavelength pairs generated from the available illumination wavelengths (Eq. 4). Surface tilt was varied using a Newport M-37S tip-tilt stage, while surface roughness was varied using roughness specimens covering multiple roughness grades. For each measurement condition, the resulting AF curves were analyzed to assess their influence on autofocus behavior and the reliability of AF-based surface reconstruction.

For the numerical processing pipeline, the axial propagation step size (Section 2.2) was set to 20 μm to provide sufficiently dense sampling of the AF response while maintaining reasonable computational cost. Gaussian spatial averaging (Section 2.3) was performed using a kernel width of 56 μm. The averaging window was chosen to encompass multiple speckles, enabling reliable estimation of local phase-gradient consistency while suppressing speckle-scale variations. The appropriate averaging scale depends on both the sample characteristics and the imaging conditions, reflecting a trade-off between AF robustness and lateral resolution.

3.1 Influence of Synthetic Wavelength

Four illumination wavelengths were combined pairwise to generate multiple synthetic wavelengths (Eq. 4), of which three representative values, approximately 494 μm, 104 μm, and 14 μm, were selected for analysis.

Figure 1 shows that the AF peak becomes progressively narrower as the synthetic wavelength decreases. Consistent with the heuristic model (Eq. 7), shorter synthetic wavelengths (corresponding to larger synthetic wavenumbers k_s) exhibit greater sensitivity to defocus, result-

ing in improved focus localization. Since the imaging system fixes both the mean wavenumber k_0 and the depth of focus (DOF $\approx$ 100 μm), synthetic wavelength selection primarily influences the AF response through k_s. Consequently, the observed sharpening of the AF response is driven largely by the choice of synthetic wavelength. Among the investigated wavelengths, $\Lambda_s = 14$ μm produced the sharpest AF response and was therefore chosen for subsequent reconstruction experiments.

While the heuristic model suggests that further reductions in synthetic wavelength could yield even sharper AF responses, arbitrarily small synthetic wavelengths are generally not preferred. Smaller synthetic wavelengths require increasingly larger separations between the illumination wavelengths (Eq. 4), which are constrained by source availability. Larger wavelength separations also increase the influence of chromatic dispersion and other wavelength-dependent effects within the measurement system.

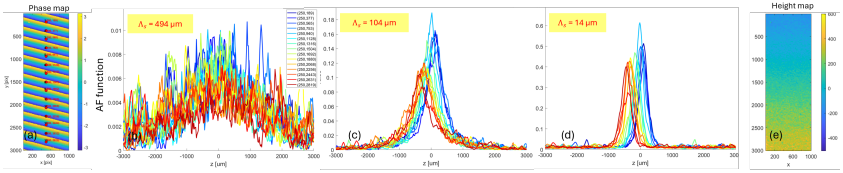


Figure 1: (a) Synthetic phase map of a tilted sample. Although the surface exhibits a continuous height variation, the phase is wrapped, making direct height estimation difficult. (b) $\Lambda_s = 494$ μm, (c) $\Lambda_s = 104$ μm, and (d) $\Lambda_s = 14$ μm. The AF curves correspond to the locations marked in (a). Decreasing the synthetic wavelength produces progressively sharper AF peaks. (e) AF height map using $\Lambda_s = 14$ μm.

3.2 Influence of Surface Tilt

Figure 2 shows that increasing surface tilt progressively sharpens the AF response. This trend is qualitatively consistent with the heuristic model presented in Section 2.4, which predicts a reduction in the roughness-induced coherence reduction term with increasing tilt. However, for the relatively small tilt angles considered here, the $\cos\alpha$

dependence in Eq. 7 alone predicts only a modest change in AF behavior.

The experimentally observed sharpening is therefore believed to be dominated by the finite numerical aperture (NA) of the imaging system. Near the zero-tilt condition, the AF response contains a significant contribution from the specular reflection, which remains relatively insensitive to defocus. As the surface tilt increases, the specular component increasingly falls outside the collection cone of the imaging system, reducing its contribution to the AF response. The remaining signal is then dominated by scattered light contributions, which decay more strongly with defocus and therefore produce sharper AF peaks.

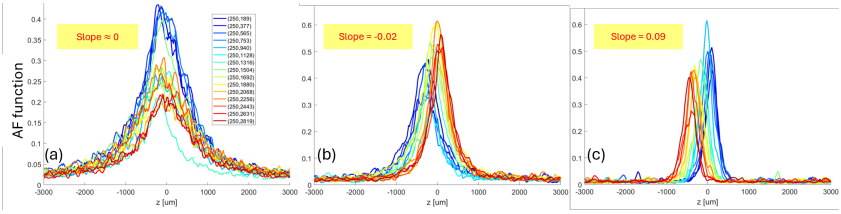


Figure 2: AF response as a function of propagation distance for three surface tilts using a fixed synthetic wavelength of $\Lambda_s = 14\ \mu$m. The curves correspond to the locations marked in Fig. 1(a). (a) Nearly flat surface ($\alpha \approx 0$), producing relatively broad AF responses. (b) Surface tilt of approximately $\alpha = -0.02$. (c) Surface tilt of approximately $\alpha = 0.09$.

3.3 Influence of Surface Roughness at Different Surface Tilts

The influence of surface roughness on the AF response is summarized in Fig. 3, where the AF width, defined as one quarter of the standard deviation of the AF response along the propagation direction, is plotted as a function of surface tilt for samples with different roughness levels. Increasing surface roughness generally results in broader AF responses, indicating reduced focus-localization capability. This effect is most pronounced near the zero-tilt condition, where the AF response is already relatively broad and therefore more sensitive to roughness-induced phase variations.

The observed behavior is consistent with the heuristic model presented in Section 2.4. According to Eq. 7, increasing surface roughness

reduces the local consistency of the phase-gradient field, resulting in a weaker and less localized AF response. As the surface tilt increases, the influence of roughness becomes less significant because the stronger phase gradients associated with the tilt contribute to a sharper AF response.

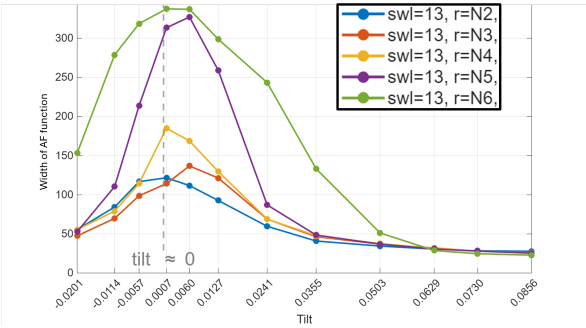


Figure 3: Dependence of AF width on surface tilt for roughness levels N2 ($R_a = 0.05\ \mu$m), N3 ($R_a = 0.10\ \mu$m), N4 ($R_a = 0.20\ \mu$m), N5 ($R_a = 0.40\ \mu$m), and N6 ($R_a = 0.80\ \mu$m), where R_a is the average absolute deviation of the surface height from the mean plane. AF width, defined as one quarter of the standard deviation of the AF response, increases with roughness, particularly near the zero-tilt condition. Small shifts in the peak locations arise from intrinsic tilt offsets of the roughness samples.

3.4 Surface Reconstruction Example

The capability of the proposed autofocus framework was evaluated on a representative curved rough surface produced by flat lapping. Figure 4 summarizes the reconstruction results obtained using a synthetic wavelength of approximately 14 μm. An amplitude-based quality mask was applied to exclude low-signal regions identified from the magnitude of the synthetic field. Signal loss near the image boundaries is attributed to the larger local surface tilt and the finite numerical aperture of the imaging system.

The resulting AF-based height map successfully recovers the overall surface topology. Despite the relatively small synthetic wavelength, height variations exceeding 600 μm were reconstructed, corresponding

to more than forty times the synthetic wavelength itself. The reconstructed surface profile is consistent with the expected geometry of the specimen and demonstrates that AF-based reconstruction is not constrained by the wrapping interval of the synthetic phase.

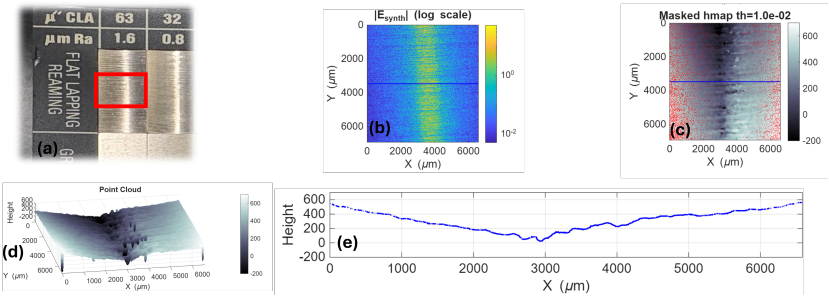


Figure 4: AF-based surface reconstruction of a curved rough sample. (a) The sample under test, with the approximate region of interest indicated by a red rectangle. (b) Magnitude of the synthetic field, used as an indicator of local signal quality. (c) Reconstructed AF height map, with low-signal regions excluded using the amplitude-based quality mask. (d) Point-cloud representation of the reconstructed surface. (e) Representative cross-section of the AF height map, showing height variations exceeding 600 μm.

Overall, the reconstruction results demonstrate that autofocus-based surface reconstruction can recover large-scale surface variations using synthetic wavelengths that are substantially smaller than the measured height range. This capability makes AF-based reconstruction a useful complement to conventional phase-based digital holography, particularly for surfaces whose height variation exceeds the unambiguous range of synthetic-phase measurements.

4 Conclusion

This work investigated AF-based surface reconstruction in dual-wavelength digital holography. Experimental results showed that the shape of the AF response depends on synthetic wavelength, surface tilt, and surface roughness. Reducing the synthetic wavelength produced progressively sharper AF peaks, improving focus localization. The heuristic model indicates that this behavior should be interpreted rela-

tive to the system DOF rather than synthetic wavelength alone. Increasing surface tilt was also found to sharpen the AF response, whereas increasing surface roughness broadened it. The observed trends were found to be in qualitative agreement with the model.

For the system investigated here, with a depth of focus of approximately 100 μm, a synthetic wavelength of $\Lambda_s = 14$ μm produced the sharpest AF response and was therefore selected for reconstruction experiments. Unlike conventional phase-based reconstruction, where larger synthetic wavelengths are generally preferred to maximize the unambiguous measurement range, AF-based reconstruction benefits from shorter synthetic wavelengths because they produce a sharper axial response. The heuristic model further suggests that arbitrarily small synthetic wavelengths are not required; rather, the synthetic wavelength should be chosen sufficiently small relative to the system DOF and the desired axial sensitivity.

Using this configuration, height variations exceeding 600 μm were successfully reconstructed on a representative rough surface. Beyond direct height reconstruction, AF-based height maps can provide the coarse height information required for hierarchical phase unwrapping. This suggests a possible route towards combining the extended measurement range of AF-based reconstruction with the high precision of phase-based measurements.

Acknowledgements

The authors thank Mitutoyo Corporation for its support and the members of the Mitutoyo Research Center Europe for their valuable discussions.

References

1. W. Gao, H. Haitjema, F. Fang, R. Leach, C. Cheung, E. Savio, and J. Linares, "On-machine and in-process surface metrology for precision manufacturing," *CIRP Annals*, vol. 68, no. 2, pp. 843–866, 2019.
2. X. Colonna De Lega, "Optical Topography Measurement of Patterned Wafers," in *AIP Conference Proceedings*, vol. 788. Richardson, Texas (USA): AIP, 2005, pp. 432–436.

3. M. Mir, B. Bhaduri, R. Wang, R. Zhu, and G. Popescu, "Quantitative Phase Imaging," in *Progress in Optics*. Elsevier, 2012, vol. 57, pp. 133–217.
4. T. Tahara, X. Quan, R. Otani, Y. Takaki, and O. Matoba, "Digital holography and its multidimensional imaging applications: A review," *Microscopy*, vol. 67, no. 2, pp. 55–67, 2018.
5. U. Schnars and W. Jueptner, *Digital Holography: Digital Hologram Recording, Numerical Reconstruction, and Related Techniques*, 1st ed. Springer Berlin, Heidelberg, 2005.
6. T. Seyler, M. Fratz, T. Beckmann, A. Bertz, and D. Carl, "Miniaturized multiwavelength digital holography sensor for extensive in-machine tool measurement," in *SPIE Optical Metrology*, P. Lehmann, W. Osten, and A. Albertazzi Gonçalves, Eds., Munich, Germany, Jun. 2017, p. 103290F.
7. S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, May 2013.
8. T. Seyler, M. Fratz, T. Beckmann, A. Schiller, A. Bertz, and D. Carl, "Extending the Depth of Field beyond Geometrical Imaging Limitations Using Phase Noise as a Focus Measure in Multiwavelength Digital Holography," *Applied Sciences*, vol. 8, no. 7, p. 1042, Jun. 2018.
9. S. Matsuura, B. Mukundakumar, R. Suzuki, and R. Barczyk, "Multiwavelength Digital Holography for Metrology: Low-Coherent Noise Sensor Configuration and Its Application," *EPJ Web of Conferences*, vol. 335, p. 01014, 2025.
10. K. Matsushima and T. Shimobaba, "Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields," *Optics Express*, vol. 17, no. 22, p. 19662, Oct. 2009.

Perspektive als 3D Messverfahren: Ein Überblick

3D Measurement using perspective: an overview

Tobias Haist, Nizar Rustom und Stephan Reichelt

Institut für Technische Optik, Universität Stuttgart
Pfaffenwaldring 9, 70569 Stuttgart

Zusammenfassung Das wichtigste 3D Messverfahren im Tierreich ist nicht etwa die Stereobildanalyse, sondern die Perspektive. In der technischen Messtechnik fristet die Perspektive allerdings eher ein Schattendasein. In den letzten Jahren haben aber bildbasierte monokulare 3D Varianten an Bedeutung gewonnen und auch die sogenannte differentielle Perspektive ist für manche Anwendungen eine interessante Variante zur 3D Positionsanalyse von Objekten im Raum. In diesem Beitrag wird auf die Grundlagen der Verfahren, vor allem in Bezug zur Triangulation, eingegangen und es werden einige Ergebnisse für einen differentiellen Perspektivensensor für die Robotik dargestellt.

Schlüsselwörter Perspektive, differentielle Perspektive, Triangulation, monokulares 3D, Größenbestimmung

Abstract The most important 3D measurement method in the animal kingdom is not stereo image analysis, but perspective. In technical measurement technology, however, perspective tends to lead a rather obscure existence. In recent years, however, image-based monocular 3D methods have gained importance, and so-called differential perspective is also an interesting alternative for some applications involving the 3D position analysis of objects in space. This paper discusses the fundamentals of these methods, particularly in relation to triangulation, and presents some results for a differential perspective sensor for robotics.

Keywords Perspektive imaging, differential perspektive, triangulation, monokular 3D

1 Einführung

Eine Vielzahl optischer Verfahren zur Messung von Entfernungen und dreidimensionalen Geometrien in allen Skalen sind verfügbar und werden in der Praxis für ganz unterschiedliche Anwendungen eingesetzt. Einteilen lassen sich die Verfahren in drei Hauptklassen:

Geometrische Verfahren: Triangulation, Perspektive, Schärfentiefebasierte Methoden, Deflektometrie

Laufzeitbasierte Verfahren: Lidar, Interferometrie, Holografie

Radiometrische Verfahren: Antastung, Shape-from-Shading, Tomografie

Wir können die geometrischen Verfahren in Gruppen und weiter in einzelne Grundmethodiken untergliedern (siehe Abb. 1). Dabei wird grafisch ersichtlich, dass letztlich bei allen diesen Methodiken Dreiecke zur Bestimmung der Entfernung genutzt werden. Aufgrund dieser Gemeinsamkeit sind auch die mathematischen Zusammenhänge ähnlich, was sich allerdings dennoch deutlich unterscheidet sind die Lage bzw. Orientierung der Dreiecke (und auch die Anzahl der Dreiecke). Entsprechend haben die Verfahren dann in der Praxis doch deutlich unterschiedliche Eigenschaften, z.B. was die Robustheit angeht.

Eine Besonderheit sind Deflektometrie (und auch Shack-Hartmann Sensorik), da dort primär nicht die Entfernung, sondern der Gradient der Entfernung erfasst wird. Das kann von Vorteil sein, aber wenn die Entfernung selbst (meist die Topografie) erfasst werden soll, dann muss der gemessene Gradient noch integriert werden (mit all den Problemen, die das mit sich bringen kann).

Alle Verfahren können als aktive oder passive Verfahren implementiert werden, also mit oder ohne explizite, strukturierte Beleuchtung. Die aktive Beleuchtung steigert in der Regel die Robustheit und Genauigkeit der Methoden.

Die Triangulation (mit all ihren Unterverfahren wie Lichtschnitt, Streifenprojektion, Stereo-Vision etc.) ist ohne Zweifel in der industriellen optischen 3D Messtechnik die mit großem Abstand wichtigste Verfahrensklasse, und auch in der Tierwelt scheint es zunächst die Methode der Wahl sein, wenn es darum geht, Entfernungen genau zu be-

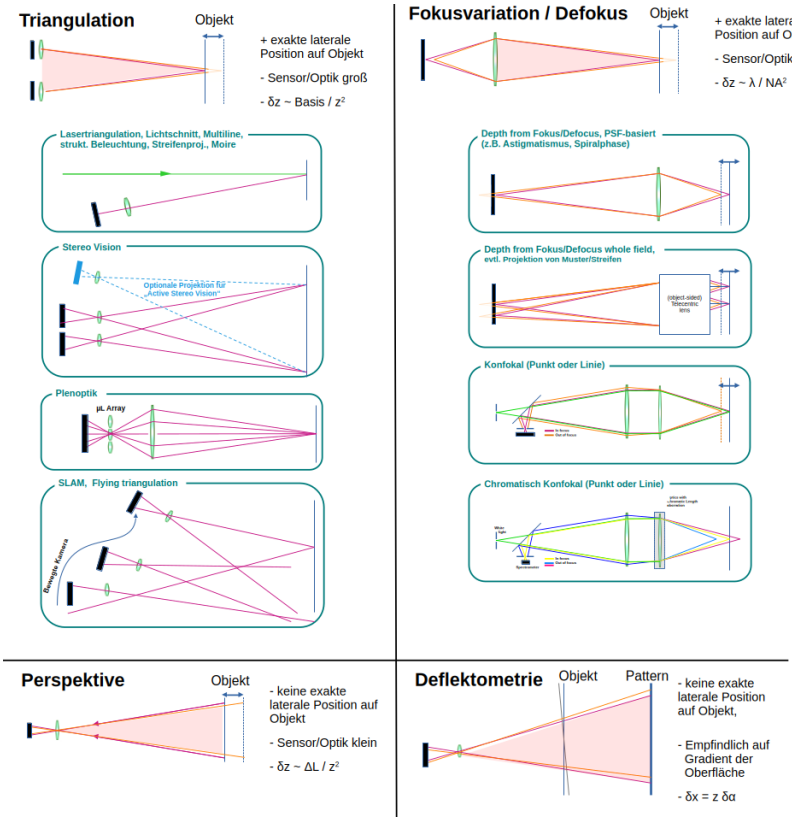


Abbildung 1: Geometrische Verfahren und — exemplarisch — ein paar der vielen Unterverfahren.

stimmen. In Wirklichkeit basiert das dreidimensionale Sehen von Tieren — egal ob Mensch oder Katze — allerdings fundamental auf der Perspektive.

In diesem Beitrag werden wir auf diese in der Technik eher stiefmütterlich behandelte Methode näher eingehen und darstellen, in welchen Formen sie sinnvoll eingesetzt werden kann und wo ihre Vorteile und Beschränkungen liegen.

2 Perspektive

Auch wenn das menschliche Sehsystem vergleichsweise schlecht darin ist, hochgenaue dreidimensionale Messungen durchzuführen, so ist es (noch) unerreicht gut darin, komplexe dreidimensionale Szenen komplett zu verstehen und zu analysieren. Dabei nutzt das Gehirn letztlich alle zur Verfügung stehenden optischen Anhaltspunkte (Verdeckungen, Schatten, Shading, Bewegungsparallaxen, Perspektive, Triangulation, Schärfentiefe, Texturgradienten) und alles verfügbare (erlernte) Vorwissen, um letztlich ein dreidimensionales Modell der Realität im Kopf entstehen zu lassen.

Wie diese innerhalb weniger hundert Millisekunden realisierte Analyse im Detail funktioniert bleibt uns ein Rätsel. Was zu welchem Grad wann genutzt wird, ist unklar. Ohne Objekte erkannt und segmentiert zu haben, ist eine Angabe hinsichtlich Verdeckungen z.B. nicht wirklich möglich. Andererseits braucht es die Verdeckungsinformation, um die Objekterkennung zu verbessern. Ein weiteres Beispiel: Der Schatten an sich wird als Schatten gedeutet, weil das zugehörige, den Schatten werfende Objekt erkannt wurde. Alles ist hier mit allem verwoben und iterativ schafft es unser Gehirn in einem holistischen Sinn, alles Wesentliche gleichzeitig zu bestimmen. Und zwar so, dass ein Modell der dreidimensionalen Realität entsteht.

Klar ist aber, dass das auch mit einem Auge, also ohne Stereovision, nahezu perfekt funktioniert (man schließe ein Auge und beobachte die Welt).

Die Basis scheint in der Tat die Kombination von Objekterkennung und Bildgrößenbestimmung y' zu sein. Bei bekannter Objektgröße y lässt sich dann gemäß der Abbildungsgleichung (bzw. letztlich dem Strahlensatz)

$$\beta' := \frac{y'}{y} = \frac{a'}{z} \approx \frac{f'}{z} \tag{1}$$

mit der Brennweite f' auf die Entfernung z zum Objekt zurückrechnen. Je besser die Objekterkennung und das Vorwissen (hinsichtlich Objektgröße y), desto besser die Entfernungsbestimmung.

Hier zeigt sich der wichtigste Unterschied zu den Triangulationsverfahren: *Es geht bei der Perspektive darum, die Entfernung zu einem mehr*

oder weniger ausgedehnten Objekt zu bestimmen, während die Triangulation die Entfernung zu einzelnen Punkten bestimmt. Dies zeigt sich in Abb. 1 darin, dass die Basis des geometrischen Bestimmungsdreiecks bei der Perspektive beim Objekt liegt und die Spitze des Dreiecks beim optischen System (im einfachsten Fall, das Loch einer Lochkamera) liegt. Bei der Perspektive und den Schärfentiefen-basierten Methoden (z.B. konfokale Mikroskopie) dreht sich dieses Verhalten um: Die Spitze des Dreiecks liegt beim Objekt (also dem Objektpunkt) und es braucht eine ausgedehnte Basis beim optischen System, sei es die Stereobasis bei der Stereovision oder auch die (numerische) Apertur bei fokusbasierten Verfahren wie z.B. der konfokalen Mikroskopie.

Anwendungsseitig bedeutet das: Wenn die Aufgabe ist, die Entfernung zu *Objekten* zu bestimmen, dann kann die Perspektive mit einem sehr kleinen, einfachen optischen System, z.B. einem Auge, das gewünschte Ergebnis liefern. Wenn stattdessen die Aufgabe die *lateral hochaufgelöste Erfassung* von Entfernungen (im Sinne z.B. einer Topografiemessung) ist, dann kann prinzipiell die Perspektive nicht mit der Triangulation konkurrieren.

Zur Analyse der erzielbaren Genauigkeit der Perspektivenmethode stellen wir Gl. 1 nach z um, also $|z| = f' y/y'$ und leiten nach der Bildgröße y', die wir bei der Messung zunächst primär bestimmen, ab und erhalten für die Unsicherheit der Entfernungsbestimmung

$$|dz| = \frac{f' y}{y'^2}\, dy' = \frac{z}{y'} dy' = \frac{z^2}{f' y}\, dy' \tag{2}$$

Wir sehen — wie übrigens bei nahezu allen geometrischen Entfernungsmessverfahren (!) — eine Zunahme der Unsicherheit der Entfernungsbestimmung mit dem Quadrat der Entfernung.

Weiterhin sehen wir, dass große Objekte zu einer geringeren Unsicherheit führen; zum einen wegen des Terms $1/y$, zum anderen, weil die Unsicherheit der Größenbestimmung des Bildes dy' umso genauer wird, je mehr (informationstragende) Pixel bei der Bildanalyse verwendet werden können.

Damit das ganze funktioniert, muss also zunächst das passende Objekt „erkannt“ werden. Im Allgemeinen eine schwierige Aufgabe, die das menschliche Sehsystem mit Bravour meistert, aber erst in den letzten Jahren in adäquater Qualität von Computer-Vision Algorithmen

beherrscht wird. Dementsprechend gewinnt auch erst in jüngerer Vergangenheit die Perspektive als Methode zur Abstandsbestimmung an Bedeutung.

Für Anwendungen, bei denen die Kamera konstant orientiert ist und die Position des Raumes bekannt ist, können auch die Entfernungen zu Objekten, die auf dem Boden stehen, rein aus der Bildgeometrie, konkret der Vertikalkoordinate im Bild bestimmt werden [1,2]. Bei Criminisi et al. wird zunächst der Fluchtpunkt einer relevanten Ebene im Bild bestimmt [2]. Für spezielle Anwendungen, z.B. Entfernung von Personen, kann man die Grundmethodik ebenfalls recht einfach anwenden [3].

3 Monokulares 3D

Für die Analyse einer Szene mit zunächst unbestimmten bzw. unbekannten Objekten können wir den einfachen Ansatz aus Gl. 1 nicht direkt ansetzen. Es müssen erst Objekte in der Szene erkannt werden und ihre reale Größe muss ermittelt bzw. geschätzt werden, um Gl. 1 anwenden zu können.

Aktuelle deep-learning basierte Verfahren der Computer-Vision vermeiden entsprechende Einzelschritte und nutzen die großen Fortschritte, die in den letzten Jahren im Bereich des Szenenverstehens erzielt wurden. Letztlich baut dabei auch der Computer (analog zum Mensch) ein Modell der Realität auf und es liegt nahe, ihn darauf zu trainieren, dass dieses Modell direkt dreidimensional entsteht. Auch wenn dabei komplette Tiefenkarten (dense depth maps) ausgegeben werden, liegt die Basis für die Bestimmung der Entfernung in einzelnen Objekten bzw. Objektteilen. (Grundsätzlich ist dabei zu unterscheiden, ob mit den Systemen „nur“ relative Aussagen oder absolute Entfernungen bestimmt werden [4].)

Wir können hier aus Platzgründen nicht auf dieses spannende neue Forschungsgebiet des monokularen 3D weiter eingehen. Eine aktuelle Übersicht hinsichtlich der sehr vielen wissenschaftlichen Arbeiten geben Zhang et al. [4,5]. Xu et al. geben einen Überblick für große Foundation Modelle mit mehr als einer Milliarde Parameter, die mit mehr als 10 Millionen Bildern trainiert werden [6]. Zum Einsatz kommen auf Transformern basierte Systeme, generative Modelle und Diffusionsmo-

delle.

Aktuell sind die Modelle erstaunlich leistungsfähig, exakte Entfernungsbestimmungen sind allerdings — wie auch für den Mensch — schwierig, so dass mit relativen Fehlern von 4 bis 5 Prozent bei Standardbenchmarks gerechnet werden muss [6]. Im Vergleich zu Systemen der optischen 3D Messtechnik natürlich ein recht schlechter Wert, der aber eben vor dem Hintergrund der minimalen Sensorik (eine Kamera) und der Komplexität der Szenen gesehen werden muss. Für die Bestimmung der relativen Anordnung von Objekten im Raum können entsprechende Systeme inzwischen in vielen Fällen mit Menschen mithalten oder diese übertreffen [7]. Danier et al. geben zumindest Indizien an, dass dabei ähnliche „visual cues“wie vom Mensch verwendet werden [8].

4 Differentielle Perspektive

Kommen wir zurück zu Gl. 1 und versuchen wir, auch ohne ausgefeilte Szenenerkennung und damit Kenntnis der Objektgröße y, die Entfernung z zu bestimmen. Die differentielle Perspektive erreicht dies, indem letztlich zwei Aufnahmen des Objekts, z.B. aus leicht unterschiedlicher Entfernung vorgenommen werden [9,10]:

$$\beta_1' = \frac{y_1'}{y} = \frac{f'}{a_1} \qquad \text{und} \qquad \beta_2' = \frac{y_2'}{y} = \frac{f'}{a_1 + \Delta a} \tag{3}$$

$$\frac{\beta_1'}{\beta_2'} = 1 + \frac{\Delta a}{a_1} \qquad \rightarrow \qquad z := a_1 = \frac{\Delta a}{y_1'/y_2' - 1} \tag{4}$$

In der Praxis ist die Methode dann bequem anwendbar, wenn wir uns selbst mit bekannter Geschwindigkeit bewegen (z.B. in einem Auto). In diesem Fall können wir eben mit einem kurzen (z.B 100 ms) Zeitversatz Δt zwei Aufnahmen machen. Da wir unsere eigene Geschwindigkeit kennen, können wir einfach das $\Delta z = \Delta a = v\Delta t$ berechnen und dann nach Gl. 4 die Entfernung zu unbewegten Objekten bestimmen.

Für eine Anwendung im Auto konnten wir mit dieser Methodik über große Entfernungen (200 m) Messunsicherheiten im Bereich 1 bis 4 m (objektabhängig) erreichen [10]. Die Perspektive ist hierbei in der Tat

eine denkbare Option, denn bei entsprechenden Anwendungen ist die Aufgabenstellung eben wirklich die Entfernungsbestimmung zu Objekten (und nicht zu einzelnen Punkten).

Wir können aber auch zu einem Einzelschussverfahren wechseln, indem wir z.B. zwei Kameras verwenden, die wir axial um Δz bzw. Δa versetzt aufstellen (Abb. 2). Hierbei sieht man gut die Verwandtschaft, aber auch den Unterschied zur Triangulation. Die laterale Separation der beiden Pupillen der Optik bei der Triangulation wird durch eine axiale Separation bei der differentiellen Perspektive ersetzt. Die Messung erfolgt axial und nicht lateral versetzt.

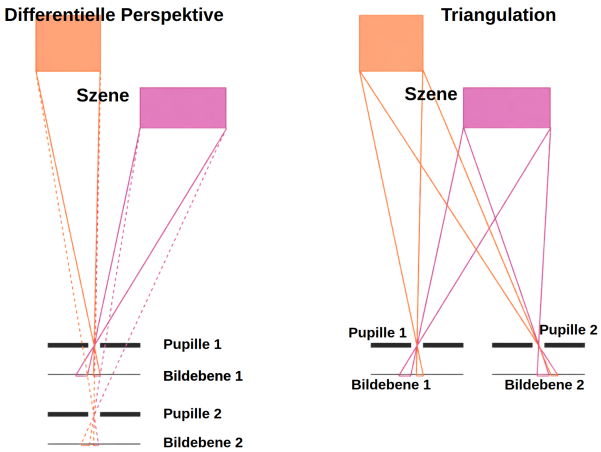


Abbildung 2: Differentielle Perspektive im Vergleich zu Triangulation: Bei der Triangulation (rechts) sind die beiden Kameras durch den Basisabstand senkrecht zur Messrichtung b getrennt. Bei der differentiellen Perspektive (links) erfolgt die Trennung in Richtung der Abstandsmessung, also koaxial. Die bei der Triangulation potentiell auftretenden Abschattungen werden so z.B. vermieden.

Abb. 3 zeigt eine realisierte Kamera basierend auf zwei preiswerten Arducam Bildsensormodulen (Sony IMX219, 3280 x 2464 Pixel, 1.12 µm Pixelgröße, Sättigungskapazität: 4300 Elektronen) mit Tamron 23FM25SP Objektiven (25 mm Brennweite, minimale Blendenzahl 1.4)

zur Abstandsbestimmung für eine Anwendung im Bereich Robotik.

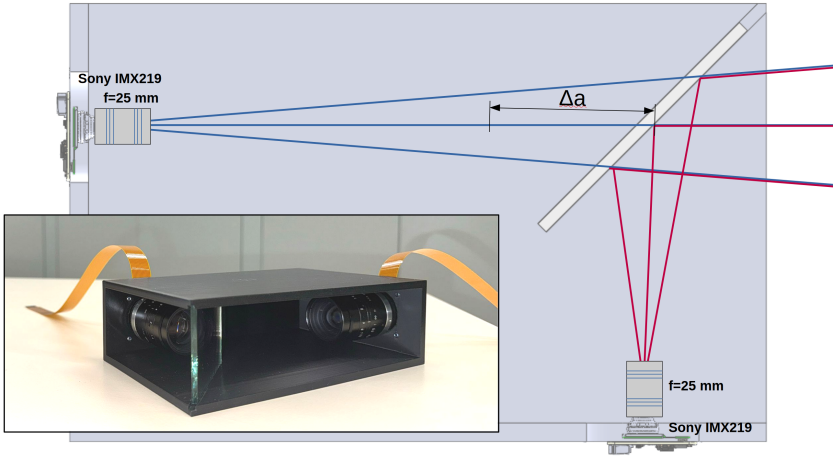


Abbildung 3: Differentielles Perspektivenmodul (ohne Front-Cover).

Verschiedene Bildverarbeitungsalgorithmen wurden eingesetzt, um letztlich das Bildgrößenverhältnis y_1'/y_2' und daraus dann den Abstand zu bestimmen. Am besten schnitt dabei eine auf Korrelation basierende Methodik ab, bei der fortlaufend beide Bilder mit unterschiedlicher Skalierung korreliert werden und dann das Maximum der Korrelation zur Bestimmung der idealen Skalierung (Bildgrößenverhältnis) genutzt wird.

Untersucht wurden die Messunsicherheit der Entfernungsbestimmung für unterschiedliche große und unterschiedlich komplexe Zufallsbinärmuster. Es zeigt sich (Abb. 4), dass — wie erwartet — zu kleine Objekte zu einer deutlich höheren Messunsicherheit führen. Die Objekte sollten auch eine gewisse Komplexität an Mustern aufweisen, allerdings (ebenfalls Abb. 4) führen zu feine Muster potentiell (je nach Objektiveinstellungen und Arbeitsbereich) zu Problemen hinsichtlich der Schärfentiefe. D.h. die feinen Muster werden am Rand des Messfeldes so unscharf, dass die Ergebnisse der Bildverarbeitung wieder

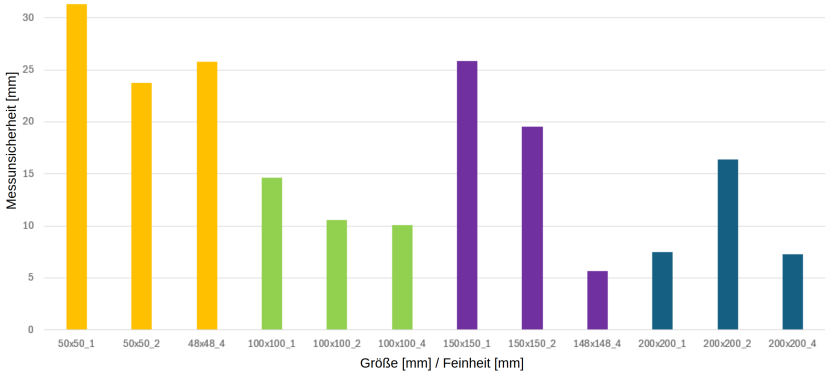


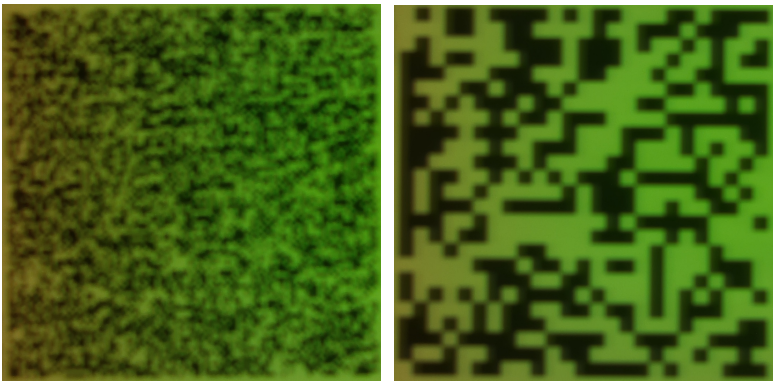
Abbildung 4: Messunsicherheiten für Entfernungsbestimmung zu QR-Codes mit dem in Abb. 3 gezeigten Modul und einer korrelationsbasierten Auswertung.

schlechter werden (siehe Abb. 5). Die dargestellten Messunsicherheiten beziehen sich auf eine Variante, bei der die Regions-of-interest vor der Korrelationsauswertung um den Faktor 4 hoch interpoliert wurden (Lanczos). Dadurch lässt sich die Genauigkeit deutlich (Faktor 4 bis 5) verbessern.

Abwandlungen der Methodik mit unterschiedlicher Brennweite findet man bei [11, 12].

5 Zusammenfassung

Die Perspektive wird als Methode zur Abstandsgewinnung in der Technik vermutlich zukünftig weiter an Bedeutung gewinnen, denn es lässt sich mit nur einem einfachen Kamerabild bei ausreichend leistungsfähiger Bildverarbeitung für viele Anwendungsszenarien — wenn auch nicht mit höchster Genauigkeit — der Abstand zu Objekten bestimmen. Darüberhinaus werden im Vergleich zur Triangulation Abschattungsproblematiken vermieden. Je besser das Szenenverstehen der Computer-Vision, desto besser wird die monokulare 3D Modellbil-



(a) Feinheit: 1 mm **(b)** Feinheit: 4 mm

Abbildung 5: Bildauschnitte für zwei Targets mit unterschiedlicher Feinheit bei Bildaufnahme am Rande des Messbereichs. Es zeigt sich, dass zu feine Details (links) aufgrund der begrenzten Schärfentiefe eben zu deutlich unscharfen Bildern und damit auch schlechteren Messunsicherheiten führen.

dung.

Die Unabhängigkeit vom Szenenverstehen sowie eine bessere Genauigkeit kann man durch die differentielle Perspektive erzielen, erkauft sich diese Vorteile aber dadurch, dass zumindest zwei Bilder zur Verfügung stehen müssen.

Vor allem muss man sich aber klar machen, dass eben mit der Perspektive *prinzipiell* nicht die Entfernung zu einem einzelnen Punkt, sondern nur zu Objekten mit einer Ausdehnung bestimmt werden kann. Denken Sie in Dreiecken!

Literatur

1. K. Koledić, L. Petrović, I. Marković, and I. Petrović, "Gvdepth: Zero-shot monocular depth estimation for ground vehicles based on probabilistic cue fusion," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025, pp. 26 126–26 135.
2. A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.
3. L. Bertoni, S. Kreiss, and A. Alahi, "Monoloco: Monocular 3d pedestrian localization and uncertainty estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6860–6870.
4. J. Zhang, Y. Wu, and H. Jiang, "Survey on monocular metric depth estimation," *Computers*, vol. 14, no. 11, p. 502, 2025.
5. Z. Zhang, Y. Zhang, Y. Li, and L. Wu, "Review of monocular depth estimation methods," *Journal of Electronic Imaging*, vol. 34, no. 2, pp. 020 901–020 901, 2025.
6. Z. Xu, H. Zhou, S. Peng, H. Lin, H. Guo, J. Shao, P. Yang, Q. Yang, S. Miao, X. He *et al.*, "Towards depth foundation models: Recent trends in vision-based depth estimation," *Computational Visual Media*, 2026.
7. D. Linsley, P. Zhou, A. Ashok, A. Nagaraj, G. S. Gaonkar, F. Lewis, Z. Pizlo, and T. Serre, "The 3d-pc: a benchmark for visual perspective taking in humans and machines," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 30 472–30 494.
8. D. Danier, M. Aygün, C. Li, H. Bilen, and O. Mac Aodha, "Depthcues: Evaluating monocular depth perception in large vision models," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 20 049–20 059.
9. A. Faulhaber, C. Krächan, and T. Haist, "Depth from axial differential perspective," *Optics Continuum*, vol. 1, no. 1, pp. 103–112, 2022.
10. T. Haist, F. Lodholz, A. Faulhaber, and S. Reichelt, "Differential perspective as a non-expensive distance sensing principle," in *Automated Visual Inspection and Machine Vision V*, vol. 12623. SPIE, 2023, pp. 91–99.
11. S. Suto, T. Matsumoto, M. Ehara, and S. Ajiki, "Distance measuring camera," Dec. 13 2022, uS Patent 11,525,676.
12. J. Ma and S. Olsen, "Depth from zooming," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 1883–1890, 1990.

Tutorial zur Hand-Auge-Kalibrierung
Tutorial on Hand-Eye Calibration

Bastian Erdnüß

Fraunhofer IOSB, Fraunhofer-Institut für
Optronik, Systemtechnik und Bildauswertung,
Fraunhoferstraße 1, 76131 Karlsruhe, Deutschland

Zusammenfassung Bei der Hand-Auge-Kalibrierung wird die Pose einer Kamera an einem Roboter-Greifarm gesucht. Der Roboter kennt die Pose seines Greifarms bzgl. seiner Basis. Die Kamerapose kann bildbasiert bzgl. eines Kalibrierkörpers oder Umgebungsmodells bestimmt werden. Die hier behandelte Hand-Auge-Kalibrierung verbindet diese beiden Welten und liefert simultan die Kamerapose am Greifarm sowie Pose und ggf. Maßstab des Kamerareferenzsystems bzgl. der Roboterbasis. Mit Ausgleichsrechnung werden sie optimal bzgl. eines vorgegebenen Fehlermodells geschätzt. Der Ansatz lässt sich auch auf die Kalibrierung anderer Starrkörperverbände übertragen.

Schlüsselwörter Hand-Auge-Kalibrierung,
Starrkörperverband, Ähnlichkeitstransformation,
Gauß-Helmert-Modell.

Abstract Hand-eye calibration determines the pose of a camera mounted on a robot arm. The robot knows the pose of its end effector relative to its base. The camera pose can be determined from images of a calibration target or a model of the environment. The hand-eye calibration presented here links these two worlds and simultaneously recovers the camera pose on the end effector as well as the pose and, where applicable, the scale of the camera reference frame relative to the robot base. A least-squares adjustment estimates them optimally with respect to a given error model. This approach also extends to the calibration of other rigid-body assemblies.

Keywords Hand-eye calibration, rigid-body assembly,
similarity transform, Gauss–Helmert model.

1 Einleitung

Industrieroboter tragen häufig eine starr am Greifarm montierte Kamera. Um aus den Bildern exakte Bewegungen ableiten zu können, muss der Roboter wissen, wie die Kamera relativ zu seinem Greifarm befestigt ist. Die Hand-Auge-Kalibrierung bestimmt diese Hand-Auge-Transformation.

Thematisch gehört die Hand-Auge-Kalibrierung in die räumliche Kalibrierung von Mehrkomponentensystemen, bei der die relativen Lagen der einzelnen Komponenten zueinander im Starrkörperverband bestimmt werden. Der einfachste Fall ist der Zweikomponentenfall, der in dieser Arbeit am Beispiel der Hand-Auge-Kalibrierung genauer betrachtet wird.

Methodisch wird die Hand-Auge-Kalibrierung hier mit den Mitteln der Ausgleichsrechnung durchgeführt. D. h., aus mehreren fehlerbehafteten Messungen wird eine bestmögliche Schätzung der Hand-Auge-Transformation ermittelt, so dass die Messfehler bzgl. der Schätzung unter Berücksichtigung des Fehlermodells minimal werden. Das Ergebnis einer solchen Schätzung hängt dabei von der Fehlermodellierung ab. Idealerweise werden dazu die tatsächlichen Messwerte modelliert, z. B. die Gelenkwinkel der Drehgeber und Featurekoordinaten im Bild, wie in [1], oder sogar direkt die Intensitätswerte des Fotosensors.

Diese ideale Modellierung ist aus praktischer Sicht jedoch oft mit schwer vertretbarem Aufwand verbunden, da so für fast jede konkrete Situation Anpassungen am Kalibrieralgorithmus erforderlich werden. Durch lineare Fehlerfortpflanzung lässt sich das Problem auf ein in erster Ordnung äquivalentes Kernproblem zurückführen, das mit einem generischen Optimierungsalgorithmus gelöst werden kann. Ein solcher Algorithmus wird in dieser Arbeit vorgestellt.

2 Problemformulierung

Zunächst werden die beteiligten Koordinatensysteme und Transformationen benannt (vgl. Abb. 1). Die Koordinatensysteme sind auf der Kameraseite (oben im Diagramm) das feste Objektsystem P und die bewegliche Kamera C, auf der Roboterseite (unten) die feste Basis B und der bewegliche Greifer G. Die beweglichen Systeme C_i und G_i (rechts)

werden dabei über mehrere Messstationen i indiziert. An jeder Station i liegen Messwerte für die Kamerapose (R_{C_i}, T_{C_i}) und die Greiferpose (R_{G_i}, T_{G_i}) vor. Die Kamerapose transformiert Kamerakoordinaten x_{C_i} in Objektkoordinaten x_P und entsprechend transformiert die Greiferpose Greiferkoordinaten x_{G_i} in Basiskoordinaten x_B:

$$x_P = R_{C_i}\, x_{C_i} + T_{C_i} \qquad \text{und} \qquad x_B = R_{G_i}\, x_{G_i} + T_{G_i}\,. \tag{1}$$

Dabei ist $R \in \mathrm{SO}(3)$ eine Rotationsmatrix und $T \in \mathbb{R}^3$ ein Translationsvektor. Statt direkter Messungen können dies auch abgeleitete Größen sein, etwa aus dem Kinematikmodell berechnete Greiferposen oder bildbasiert durch Rückwärtsschnitt bestimmte Kameraposen.

Von Objektsystem P zu Roboterbasis B wird eine Ähnlichkeitstransformation $(S\,R_P, T_P)$ mit Maßstab $S \neq 0$ angesetzt. Von Kamera C_i zu Greifer G_i besteht dann ebenfalls eine Ähnlichkeitstransformation $(S\,R_Q, T_Q)$ mit demselben Maßstab S. Diese ist für alle Stationen i gleich, da die Kamera starr mit dem Greifer verbunden ist. (R_P, T_P) ist die Objektpose in der Roboterbasis und (R_Q, T_Q) die gesuchte Hand-Auge-Transformation. Für die Koordinaten der beteiligten Systeme gilt:

$$x_B = S\,R_P\, x_P + T_P \qquad \text{und} \qquad x_{G_i} = S\,R_Q\, x_{C_i} + T_Q\,. \tag{2}$$

Häufig ist der Maßstab S unbekannt und muss mitberücksichtigt werden, da eine Kamera den Maßstab ihrer Umgebung nicht direkt erfassen kann. Wurden die Kameraposen allerdings gegen ein Kalibrierobjekt bestimmt, kann der Maßstab S aus dessen Objektgröße bekannt sein, was hier als Sonderfall mitbehandelt wird.

Die beiden Pfade $B \leftarrow P \leftarrow C_i$ und $B \leftarrow G_i \leftarrow C_i$ müssen übereinstimmen (vgl. Abb. 1). Eliminiert man x_P und x_{G_i} aus (1) und (2)

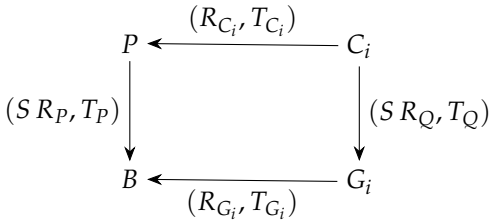


Abbildung 1: Beteiligte Koordinatensysteme und Transformationen.

und setzt die beiden entstehenden Ausdrücke für x_B gleich, liefert Koeffizientenvergleich an x_{C_i} und den konstanten Termen die Gleichungen

$$R_{G_i} R_Q = R_P R_{C_i} \qquad \text{und} \qquad T_P - R_{G_i} T_Q + S R_P T_{C_i} = T_{G_i} . \quad (3)$$

Dies sind zwei über R_P gekoppelte Gleichungssysteme, die für die tatsächlichen Posen exakt gelten. Gesucht sind daher (R_P, T_P), (R_Q, T_Q) und ggf. S gemeinsam mit Kamera- und Greiferposen, die (3) exakt erfüllen und mit den gemessenen Posen im Rahmen ihrer Genauigkeit übereinstimmen.

3 Lösungsmethode

Zu vorgegebenem R_P ist die rechte Gleichung in (3) linear in T_P, T_Q und S und kann als überbestimmtes Gleichungssystem mit der Methode der kleinsten Fehlerquadrate nach T_P, T_Q und ggf. S gelöst werden. Für die linke Gleichung in (3) wurde in der Literatur eine Vielzahl von Lösungsmöglichkeiten vorgeschlagen, über die Shah et al. 2012 in [2] einen umfassenden Überblick gaben. Ein Jahr später präsentierte Shah in [3] eine konzeptionell sehr einfache Methode. Dabei wird die 9×9-Hilfsmatrix $K = \sum_i R_{C_i} \otimes R_{G_i}$ mit dem Kroneckerprodukt $\otimes$ gebildet und zu $U \Sigma V^\top = K$ singulärwertzerlegt. Aus den singulären Vektoren u_0 und v_0 zum größten singulären Wert σ_0 ergeben sich R_P und R_Q durch spaltenweises Umformen der 9D-Vektoren in 3×3-Matrizen mit anschließender Vorzeichennormierung und Projektion auf $\mathrm{SO}(3)$.

Für dieses Verfahren benötigt man keine initialen Schätzungen der Parameter. Messfehler wirken sich jedoch in nicht kontrollierbarer Weise auf das Ergebnis aus. Dennoch bekommt man mit dem Verfahren eine Lösung, die als Startwert für eine iterative Optimierung verwendet werden kann. Das motiviert den folgenden Initialisierungsalgorithmus:

```
1 import numpy as np
2 from scipy.spatial.transform import Rotation as Rot
3 Id = np.eye(3)
4
5 S = None # oder auf bekannten Massstab setzen
6 # Lade Messdaten: RC, TC, RG, TG, n
```

```
assert RC.shape == RG.shape == (n, 3, 3)
assert TC.shape == TG.shape == (n, 3, 1)

K = sum(np.kron(rc, rg) for rc, rg in zip(RC, RG))
U, Sigma, Vt = np.linalg.svd(K)
RP, RQ = U[:,0].reshape(3, 3).T, Vt[0,:].reshape(3, 3).T
# Vorzeichennormierung und Projektion auf SO(3):
RP = Rot.from_matrix(RP * np.linalg.det(RP)).as_matrix()
RQ = Rot.from_matrix(RQ * np.linalg.det(RQ)).as_matrix()

if S is None:
    A = [[Id, -rg, RP @ tc] for rg, tc in zip(RG, TC)]
    b = TG
else:
    A = [[Id, -rg] for rg in RG]
    b = [tg - S * RP @ tc for tg, tc in zip(TG, TC)]

A, b = np.block(A), np.vstack(b)
x, _, _, sv = np.linalg.lstsq(A, b, rcond=None)
TP, TQ = x[0:3], x[3:6]
if S is None: S = x[6,0]
```

Für eine eindeutige Lösung sind mindestens drei Messungen ausreichend verschiedener Posen erforderlich. Gilt nicht `Sigma[0]` $\approx n$, sind die Rotationsmessungen widersprüchlich. Ist `Sigma[1]` $\approx$ `Sigma[0]`, haben die Rotationen eine nahezu gemeinsame Achse und R_P, R_Q wurden nicht eindeutig bestimmt. Ist `sv[-1]` ≈ 0, gilt dies für T_P, T_Q, S.

3.1 Einschub: Ausgleichsrechnung

Im Gauß-Helmert-Modell der Ausgleichsrechnung kann eine optimale Lösung durch explizite Modellierung der Messfehler bestimmt werden (vgl. [4, Kap. 4.8]). Dazu werden mit Kovarianzmatrix Σ zentriert normalverteilte Residuen v an die Messwerte eingeführt, die zu den Messwerten addiert die tatsächlichen Werte ergeben. Sie entsprechen im Wesentlichen den Messfehlern, nur dass sie an den Messwerten anliegen, anstatt an den unbekannten wahren Werten. Zusätzlich werden Zwangsbedingungen $h(x, v) = 0$ zwischen den Parametern x und den Residuen v formuliert.

Ausgehend von einer aktuellen Schätzung (x, v) wird eine verbesserte Schätzung $(\widehat{x}, \widehat{v}) = (x + \Delta x, v + \Delta v)$ bestimmt, die die an (x, v) linearisierte Zwangsbedingung

$$h(\widehat{x}, \widehat{v}) \approx h + A\,\Delta x + B\,(\widehat{v} - v) = 0 \tag{4}$$

mit der Designmatrix $A = \mathrm{D}_x h(x, v)$, der Bedingungsmatrix $B = \mathrm{D}_v h(x, v)$ und dem A-priori-Widerspruch $h = h(x, v)$ erfüllt und die Lagrange-Funktion

$$L = \tfrac{1}{2}\|\widehat{v}\|^2_{\Sigma^{-1}} - \lambda^\top (h + A\,\Delta x + B\,(\widehat{v} - v)) \tag{5}$$

mit den Lagrange-Faktoren λ minimiert. Mit dem Ansatz $\widehat{v} = \Sigma\, B^\top \lambda + a$ kann durch Ausmultiplizieren der quadratischen Norm $\|\widehat{v}\|^2_{\Sigma^{-1}} = \widehat{v}^\top \Sigma^{-1}\, \widehat{v}$ nach Zusammenfassen mit dem hinteren $\widehat{v}$-Term geprüft werden, dass L für $a = 0$ minimal wird. Setzt man daher $\widehat{v} = \Sigma\, B^\top \lambda$ in (4) ein, folgt $B\,\Sigma\, B^\top \lambda = B\,v - h - A\,\Delta x$, was nach λ gelöst werden kann. Das resultierende λ zusammen mit $\widehat{v}$ in (5) eingesetzt, liefert das Kleinste-Quadrate-Funktional für das i. d. R. überbestimmte Gleichungssystem $A\,\Delta x = B\,v - h$ in Δx bzgl. der Kovarianzmatrix $B\,\Sigma\, B^\top$. Das lässt sich zu den Lösungsschritten

$$A\,\Delta x = B\,v - h \quad \sim \quad B\,\Sigma\, B^\top \qquad \Longrightarrow\ \Delta x \tag{6}$$

$$B\,\Sigma\, B^\top \lambda = B\,v - h - A\,\Delta x \qquad \Longrightarrow\ \lambda \tag{7}$$

$$\widehat{v} = \Sigma\, B^\top \lambda \qquad \Longrightarrow\ \widehat{v} \tag{8}$$

zusammenfassen.

3.2 Einschub: Lie-Algebra der Rotationsgruppe

Die Rotationsmatrizen bilden zusammen mit der Matrixmultiplikation die Lie-Gruppe SO(3) mit der Lie-Algebra $\mathfrak{so}(3)$ der schiefsymmetrischen 3×3-Matrizen, die isomorph zu den Rotationsvektoren $\mathbb{R}^3$ mit dem Kreuzprodukt $\times$ als Multiplikation ist. $[r]_\times$ bezeichnet die schiefsymmetrische Matrix, für die $[r]_\times\, t = r \times t$ für alle t gilt.

Eine Lie-Gruppe ist eine Mannigfaltigkeit und die zugehörige Lie-Algebra ist auch gleichzeitig deren Tangentialraum am Einselement, welches für SO(3) die Einheitsmatrix I als Null-Rotation ist. Durch

geodätisches Fortschreiten von I entlang des Rotationsvektors $r \in \mathbb{R}^3$ wird die Rotationsmatrix

$$\mathrm{E}^r := \mathrm{Exp}(r) := \exp([r]_\times) \tag{9}$$

erreicht, wobei exp die gewöhnliche Matrixexponentialfunktion ist.

Solà et al. geben in [5] eine ausführliche Darstellung der hier benötigten Lie-Theorie, insb. Example 3 und 4 sowie Abschnitt II.D bis II.F und Appendix B. In dieser Arbeit werden die linksseitigen Operationen [5, (27) und (28)] genutzt.

Die folgenden weiteren Rechenregeln für Exp werden benötigt:

$$\begin{aligned} &(\mathrm{E}^r)^{-1} = (\mathrm{E}^r)^\top = \mathrm{E}^{-r}, \quad R\,\mathrm{E}^r = \mathrm{E}^{R\,r}\,R, \quad J_l(-r) = J_l(r)^\top, \\ &\mathrm{E}^{\Delta r_1}\mathrm{E}^{\Delta r_2} \approx \mathrm{E}^{\Delta r_1 + \Delta r_2}, \qquad \mathrm{E}^{r+\Delta r} \approx \mathrm{E}^{J_l(r)\,\Delta r}\,\mathrm{E}^r, \end{aligned} \tag{10}$$

wobei die letzten beiden für kleine $\|\Delta r_1\|, \|\Delta r_2\|$ und $\|\Delta r\|$ in erster Ordnung gelten. $J_l(r)$ ist dabei mit $\mathrm{sinc}(a) = \sin(a)/a$ und $a = \|r\|/2$ die linke Jacobi-Matrix

$$J_l(r) := I + \frac{\mathrm{sinc}^2(a)}{2}\,[r]_\times + \frac{1 - \mathrm{sinc}(a)\,\cos(a)}{\|r\|^2}\,[r]_\times^2 \tag{11}$$

der Exponentialabbildung Exp, vgl. [5, (145)]. (11) ist für kleine $\|r\|$ numerisch unproblematisch, $J_l(0) = I$ muss jedoch gesondert berechnet werden.

Mit der zweiten Formel in (10) werden kleine Rotationsupdates in Rotationsketten algebraisch nach links sortiert. Mit den letzten beiden können kleine Rotationsupdates zusammengefasst, oder kleine Fehlerupdates von der bisherigen Fehlerannahme separiert werden. Eine typische Rechnung kann daher wie folgt aussehen:

$$\begin{aligned} &(\mathrm{E}^{-\Delta r_0}\,R\,\mathrm{E}^{r+\Delta r})^\top = \mathrm{E}^{-r-\Delta r}\,R^\top\,\mathrm{E}^{\Delta r_0} \approx \mathrm{E}^{-J_l(-r)\,\Delta r}\,\mathrm{E}^{-r}\,R^\top\,\mathrm{E}^{\Delta r_0} \\ &= \mathrm{E}^{-J_l(r)^\top\,\Delta r}\,\mathrm{E}^{\mathrm{E}^{-r}\,R^\top\,\Delta r_0}\,\mathrm{E}^{-r}\,R^\top = \mathrm{E}^{\mathrm{E}^{-r}\,R^\top\,\Delta r_0 - J_l(r)^\top\,\Delta r}\,\mathrm{E}^{-r}\,R^\top. \end{aligned} \tag{12}$$

3.3 Optimale Lösung

Um Abschnitt 3.1 auf die Hand-Auge-Kalibrierung anwenden zu können, muss das Problem erst in die richtige Form gebracht werden.

Die Zwangsbedingungen ergeben sich aus (3) und werden in Bedingungen f_i für die Rotationen und g_i für die Translationen aufgeteilt. Für die Translationen geht das etwas einfacher, für sie kann einfach die Differenz

$$g_i := T_P - R_{G_i} T_Q + S R_P T_{C_i} - T_{G_i} \tag{13}$$

genommen werden. Das Äquivalent für die Rotationen ist

$$f_i := \operatorname{Log}\left(R_P R_{C_i} (R_{G_i} R_Q)^\top\right), \tag{14}$$

was dem Rotationsvektor f_i entspricht, für den $\mathrm{E}^{f_i} = R_P R_{C_i} (R_{G_i} R_Q)^\top$ gilt. Für $f_i = 0$ ist $\mathrm{E}^{f_i} = I$, also $R_P R_{C_i} = R_{G_i} R_Q$ die Bedingung aus (3). f_i und g_i werden zu einer Stationsbedingung $h_i = (f_i, g_i)$ gestapelt und alle Stationsbedingungen zu einer Gesamtbedingung $h = (h_1, \ldots, h_n)$.

Die Parameter $x = (R_P, T_P, R_Q, T_Q, S)$ leben auf einer Mannigfaltigkeit und die Updates

$$\Delta x = (\Delta r_P, \Delta t_P, \Delta r_Q, \Delta t_Q, \Delta s) \tag{15}$$

in ihrem Tangentialraum. Die Addition $\widehat{x} = x + \Delta x$ wird darauf als

$$\begin{aligned} \widehat{R}_P &= \mathrm{E}^{\Delta r_P} R_P & \widehat{T}_P &= T_P + \Delta t_P & \widehat{S} &= \mathrm{e}^{\Delta s} S \\ \widehat{R}_Q &= \mathrm{E}^{\Delta r_Q} R_Q & \widehat{T}_Q &= T_Q + \Delta t_Q \end{aligned} \tag{16}$$

verstanden. Die Rotationen werden dabei gemäß Abschnitt 3.2 linksseitig aktualisiert und der Maßstab S multiplikativ mit der Exponentialfunktion $\mathrm{e}^{\Delta s}$. Durch letzteres ist sichergestellt, dass $\widehat{S}$ das Vorzeichen nicht wechseln kann und auch nicht verschwindet.

Da h aus den Stationsbedingungen besteht, besteht auch die Designmatrix $A = \mathrm{D}_x h$ aus den gestapelten Stationsdesignmatrizen $A_i = \mathrm{D}_x h_i$, die ihrerseits wieder aus den beiden gestapelten Untermatrizen $\mathrm{D}_x f_i$ und $\mathrm{D}_x g_i$ bestehen. Um diese zu berechnen, werden f_i und g_i an $\widehat{x}$ betrachtet:

$$\mathrm{E}^{\widehat{f_i}} = \widehat{R}_P R_{C_i} \widehat{R}_Q^\top R_{G_i}^\top = \mathrm{E}^{\Delta r_P} R_P R_{C_i} R_Q^\top \mathrm{E}^{-\Delta r_Q} R_{G_i}^\top \tag{17}$$

$$\begin{aligned} \widehat{g}_i &= \widehat{T}_P - R_{G_i} \widehat{T}_Q + \widehat{S} \widehat{R}_P T_{C_i} - T_{G_i} \\ &= T_P + \Delta t_P - R_{G_i} (T_Q + \Delta t_Q) + \mathrm{e}^{\Delta s} S \mathrm{E}^{\Delta r_P} R_P T_{C_i} - T_{G_i}. \end{aligned} \tag{18}$$

Mit (10) wird in (17) analog zu (12) erst Δr_Q nach links sortiert und mit Δr_P zusammengefasst. Rechts bleibt E^{f_i}, was für kleine f_i mit zusammengefasst werden kann. Exponentenvergleich liefert dann

$$\widehat{f_i} \approx f_i + \mathrm{D}_x f_i \, \Delta x = f_i + \Delta r_P - R_P \, R_{C_i} \, R_Q^\top \, \Delta r_Q \, . \tag{19}$$

Aus (18) folgt mit den Näherungen $\mathrm{e}^{\Delta s} \approx 1 + \Delta s$, $\mathrm{E}^{\Delta r_P} \approx I + [\Delta r_P]_\times$ und $\Delta s \, \Delta r_P \approx 0$ für kleine Δs und Δr_P und der Abkürzung $T_i = S \, R_P \, T_{C_i}$

$$\widehat{g_i} \approx g_i + \mathrm{D}_x g_i \, \Delta x = g_i + \Delta t_P - R_{G_i} \, \Delta t_Q + \Delta s \, T_i - [T_i]_\times \, \Delta r_P \, , \tag{20}$$

wobei im letzten Schritt mit $\Delta r_P \times T_i = -T_i \times \Delta r_P$ die Antikommutativität des Kreuzprodukts genutzt wurde. Um $\Delta r_P, \Delta r_Q$ und f_i zusammenfassen zu können und damit die Näherungen in Δs und Δr_P gelten, müssen diese Größen klein sein. Das trifft erst gegen Ende der Iteration zu, wenn sich die Parameter nur noch minimal ändern und die Zwangsbedingungen praktisch erfüllt sind. Dort entsteht dann die korrekte Gleichgewichtsbedingung für eine optimale Lösung. Bei großen Fehlern kann in den ersten Schritten eine schnellere Konvergenz erreicht werden, wenn stattdessen $\mathrm{E}^{\Delta f_i} \, \mathrm{E}^{f_i} \approx \mathrm{E}^{f_i + J_l(f_i)^{-1} \, \Delta f_i}$ zusammengefasst wird.

Um die Residuen zu modellieren, sollen von nun an (R_{C_i}, T_{C_i}) und (R_{G_i}, T_{G_i}) die optimal geschätzten Posen bezeichnen (oder zumindest die aktuell beste Approximation daran), wohingegen die gemessenen Posen (R_{c_i}, T_{c_i}) und (R_{g_i}, T_{g_i}) zur Unterscheidung mit kleinen Buchstaben im Index geschrieben werden. Die Residuen

$$v_i = (r_{C_i}, t_{C_i}, r_{G_i}, t_{G_i}) \tag{21}$$

sind Tangentialvektoren an die Messwerte, so dass für die geschätzten Posen

$$\begin{aligned} R_{C_i} &= \mathrm{E}^{r_{C_i}} \, R_{c_i} & \qquad T_{C_i} &= T_{c_i} + t_{C_i} \\ R_{G_i} &= \mathrm{E}^{r_{G_i}} \, R_{g_i} & \qquad T_{G_i} &= T_{g_i} + t_{G_i} \end{aligned} \tag{22}$$

gilt. Die Stationsresiduen v_i werden zu einem Gesamtresiduum $v = (v_1, \ldots, v_n)$ zusammengefasst. Die Stationsbedingungen (13) und (14) gelten für die geschätzten Posen und hängen mit (22) in (13) und (14) eingesetzt von den gemessenen Posen und den Residuen ab. Da

die Stationsbedingung h_i nur von ihren Stationsresiduen v_i abhängt, wird die Bedingungsmatrix B zu einer Blockdiagonalmatrix mit den rechteckigen Diagonalblöcken $B_i = \mathrm{D}_{v_i} h_i$, die aus den beiden übereinandergestapelten Untermatrizen $\mathrm{D}_{v_i} f_i$ und $\mathrm{D}_{v_i} g_i$ bestehen. Um diese zu berechnen, werden f_i und g_i an $\widehat{v}_i = v_i + \Delta v_i$ betrachtet. Mit (22) eingesetzt gilt für diese

$$\mathrm{E}^{\widehat{f}_i} = R_P \, \mathrm{E}^{\widehat{r}_{C_i}} \, R_{c_i} \, R_Q^\top \, R_{g_i}^\top \, \mathrm{E}^{-\widehat{r}_{G_i}} \tag{23}$$

$$\widehat{g}_i = T_P - \mathrm{E}^{\widehat{r}_{G_i}} \, R_{g_i} \, T_Q + S \, R_P \, (T_{c_i} + \widehat{t}_{C_i}) - T_{g_i} - \widehat{t}_{G_i} \, . \tag{24}$$

$\widehat{r}_{C_i} = r_{C_i} + \Delta r_{C_i}$ und $\widehat{r}_{G_i} = r_{G_i} + \Delta r_{G_i}$ in (23) eingesetzt und mit (10) wie in (12) die Δr-Änderungen mit der Links-Jacobi-Matrix von den Fehlern separiert und nach links sortiert, lässt rechts E^{f_i} übrig. Alles zusammengefasst ergibt nach Koeffizientenvergleich

$$\begin{aligned} \widehat{f}_i &\approx f_i + \mathrm{D}_{v_i} f_i \, \Delta v_i \\ &= f_i + R_P \, J_l(r_{C_i}) \, \Delta r_{C_i} - R_P \, \mathrm{E}^{r_{C_i}} \, R_{c_i} \, R_Q^\top \, R_{g_i}^\top \, J_l(r_{G_i})^\top \, \Delta r_{G_i} \, . \end{aligned} \tag{25}$$

$\widehat{t}_{C_i} = t_{C_i} + \Delta t_{C_i}$, $\widehat{t}_{G_i} = t_{G_i} + \Delta t_{G_i}$ und $\widehat{r}_{G_i} = r_{G_i} + \Delta r_{G_i}$ in (24) eingesetzt ergeben mit (10) für Δr_{G_i} und der Näherung $\mathrm{E}^r \approx I + [r]_\times$ für kleine r

$$\begin{aligned} \widehat{g}_i &\approx g_i + \mathrm{D}_{v_i} g_i \, \Delta v_i \\ &= g_i + [\mathrm{E}^{r_{G_i}} \, R_{g_i} \, T_Q]_\times \, J_l(r_{G_i}) \, \Delta r_{G_i} + S \, R_P \, \Delta t_{C_i} - \Delta t_{G_i} \, , \end{aligned} \tag{26}$$

wobei auch wieder die Antikommutativität des Kreuzprodukts genutzt wurde.

Mit (22) in (19) und (20) eingesetzt, lassen sich zusammen mit (25) und (26) unter Berücksichtigung der Reihenfolgen in (15) und (21) die Stationsdesign- und -bedingungsmatrixblöcke

$$A_i = \begin{pmatrix} I & 0 & -R_P \, \mathrm{E}^{r_{C_i}} \, R_{c_i} \, R_Q^\top & 0 & 0 \\ -[T_i]_\times & I & 0 & -\mathrm{E}^{r_{G_i}} \, R_{g_i} & T_i \end{pmatrix} \tag{27}$$

$$B_i = \begin{pmatrix} R_P \, J_l(r_{C_i}) & 0 & -R_P \, \mathrm{E}^{r_{C_i}} \, R_{c_i} \, R_Q^\top \, R_{g_i}^\top \, J_l(r_{G_i})^\top & 0 \\ 0 & S \, R_P & [\mathrm{E}^{r_{G_i}} \, R_{g_i} \, T_Q]_\times \, J_l(r_{G_i}) & -I \end{pmatrix} \tag{28}$$

mit $T_i = S \, R_P \, (T_{c_i} + t_{C_i})$ ablesen. Bei bekanntem Maßstab S muss nur die letzte Spalte von A_i gestrichen werden und Δx enthält kein Δs mehr.

3.4 Kovarianzmatrizen

Die Stationsresiduen v_i können korreliert sein. Fehler in der Kamerakalibrierung können sich z. B. in Form von systematischen Fehlern in den Posenschätzungen auswirken, wodurch diese korreliert sind. Häufig sind kleine Korrelationen jedoch vernachlässigbar. Für unkorrelierte Stationsresiduen ist die Kovarianzmatrix Σ der Residuen v eine Blockdiagonalmatrix mit Diagonalblöcken Σ_i. Durch die Blockdiagonalstruktur von B wird dann auch $B\,\Sigma\,B^\top$ blockdiagonal mit den Diagonalblöcken $B_i\,\Sigma_i\,B_i^\top$, und (7) und (8) lassen sich blockweise lösen. (6) ist meist hinreichend gut konditioniert, so dass es sich über die Normalgleichungen lösen lässt. Insgesamt ergeben sich daher mit A_i und B_i aus (27) und (28) die vereinfachten Lösungsschritte

$$W_x = \sum_i A_i^\top \,(B_i\,\Sigma_i\,B_i^\top)^{-1}\,A_i \tag{29}$$

$$\Delta x = W_x^{-1} \sum_i A_i^\top \,(B_i\,\Sigma_i\,B_i^\top)^{-1}\,(B_i\,v_i - h_i) \tag{30}$$

$$\widehat{v}_i = \Sigma_i\,B_i^\top \,(B_i\,\Sigma_i\,B_i^\top)^{-1}\,(B_i\,v_i - h_i - A_i\,\Delta x)\,. \tag{31}$$

Ausgehend von $(x, v) = (x_0, 0)$ mit der Initialschätzung x_0 aus Abschnitt 3 kann iterativ Δx nach (30) bestimmt werden und daraus $\widehat{v}$ nach (31). Zusammen mit $\widehat{x} = x + \Delta x$ bildet $(x, v) = (\widehat{x}, \widehat{v})$ den Startwert für die nächste Iteration. Dies wird wiederholt, bis $\|\Delta x\|^2_{W_x}$ hinreichend klein ist. Am Ende minimiert x lokal die Residuenquadratsumme $\sum_i \|v_i\|^2_{\Sigma_i^{-1}}$ bzgl. der Kovarianzen Σ_i. Die optimalen Posen folgen bei Bedarf aus (22).

Werden die Roboter- und Kameraposen mit unabhängigen Verfahren bestimmt, sind sie zwangsläufig unkorreliert und Σ_i wird damit selbst blockdiagonal. Wegen der Wahl der Reihenfolge in (21) ist

$$\Sigma_i = \begin{pmatrix} \Sigma_{C_i} & 0 \\ 0 & \Sigma_{G_i} \end{pmatrix} \quad \Sigma_{C_i} = \begin{pmatrix} \Sigma_{r_{C_i}} & V_{C_i}^\top \\ V_{C_i} & \Sigma_{t_{C_i}} \end{pmatrix} \quad \Sigma_{G_i} = \begin{pmatrix} \Sigma_{r_{G_i}} & V_{G_i}^\top \\ V_{G_i} & \Sigma_{t_{G_i}} \end{pmatrix}, \tag{32}$$

wobei V_{C_i} die Kovarianz von t_{C_i} und r_{C_i} ist und V_{G_i} analog für G_i. Beim Roboter kann Σ_{G_i} per Vorwärtskinematik aus der Genauigkeit der als unkorreliert angenommenen Gelenkwinkel der Drehgeber berechnet werden. Eine ideale Pan-Tilt-Einheit mit Azimut a_i und Elevation b_i

im Bogenmaß und Residuen α_i und β_i liefert mit den Einheitsvektoren $e_1 = (1,0,0)$ und $e_3 = (0,0,1)$ z. B. die Posen $(R_{g_i}, T_{g_i}) = (\mathrm{E}^{a_i\, e_3}\, \mathrm{E}^{b_i\, e_1}, 0)$ mit Residuen $(r_{G_i}, t_{G_i}) = (\alpha_i\, e_3 + \beta_i\, \mathrm{E}^{a_i\, e_3}\, e_1, 0)$, da $\mathrm{E}^{(a_i+\alpha_i)\, e_3}\, \mathrm{E}^{(b_i+\beta_i)\, e_1} \approx \mathrm{E}^{r_{G_i}}\, R_{g_i}$ ist. Mit $J_i = (e_3, \mathrm{E}^{a_i\, e_3}\, e_1)$ sind dann $\Sigma_{t_{G_i}} = V_{G_i} = 0$ und

$$\Sigma_{r_{G_i}} = J_i\, \mathrm{diag}(\sigma_\alpha^2, \sigma_\beta^2)\, J_i^\top = \sigma_\alpha^2\, e_3\, e_3^\top + \sigma_\beta^2\, (\mathrm{E}^{a_i\, e_3}\, e_1)\, (\mathrm{E}^{a_i\, e_3}\, e_1)^\top \quad (33)$$

mit den Standardabweichungen σ_α und σ_β von α_i und β_i.

Stammen die Kameraposen aus einem Bündelblockausgleich, kann dieser die geschätzten Kovarianzmatrizen Σ_{C_i} liefern. Dabei ist die vom Bündelblock gewählte Konvention für die Angabe der Kovarianzmatrix zu beachten und ggf. auf die hier genutzte Konvention anzupassen, bei der Posen (R, T) raumfest $C \to P$ angegeben werden, von links auf $\mathrm{SO}(3) \times \mathbb{R}^3$ zu $(\mathrm{E}^r\, R, T + t)$ gestört werden und in der Reihenfolge (r, t) in der Kovarianzmatrix stehen.

Stehen keine Kovarianzmatrizen zur Verfügung, kann mit isotropen Fehlern $\Sigma_{r_{C_i}} / \sigma_{r_C}^2 = \Sigma_{t_{C_i}} / \sigma_{t_C}^2 = \Sigma_{r_{G_i}} / \sigma_{r_G}^2 = \Sigma_{t_{G_i}} / \sigma_{t_G}^2 = I$ angesetzt werden und $\sigma_{r_C}^2, \sigma_{t_C}^2, \sigma_{r_G}^2$ und $\sigma_{t_G}^2$ per Varianzkomponentenschätzung aus den Daten bestimmt werden, vgl. [4, Kap. 4.2.4]. Dafür sind ausreichend viele Messungen erforderlich.

Literatur

1. M. Ulrich, C. Steger, F. Butsch, and M. Liebe, "Vision-guided robot calibration using photogrammetric methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 218, pp. 645–662, 2024.
2. M. Shah, R. D. Eastman, and T. Hong, "An overview of robot-sensor calibration methods for evaluation of perception systems," in *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, ser. PerMIS '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 15–20.
3. M. Shah, "Solving the robot-world/hand-eye calibration problem using the Kronecker product," *Journal of Mechanisms and Robotics*, vol. 5, no. 3, June 2013.
4. W. Förstner and B. P. Wrobel, *Photogrammetric computer vision*, ser. Geometry and Computing. Springer Cham, 2016.
5. J. Solà, J. Deray, and D. Atchuthan, "A micro Lie theory for state estimation in robotics," 2021. [Online]. Available: https://arxiv.org/abs/1812.01537

A brief overview of quantum and quantum-inspired approaches for image segmentation

Marie Gogolin[1,3], Yannick Werner[2,3], Alexander Geng[1], Paul Lukowicz[2,3], Ali Moghiseh[1], and Maximilian Kiefer-Emmanouilidis[2,3]

[1] Fraunhofer Institute for Industrial Mathematics ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
[2] DFKI German Research Center for Artificial Intelligence, Trippstadter Str. 122, 67663 Kaiserslautern, Germany
[3] RPTU Kaiserslautern-Landau, Gottlieb-Daimler-Straße, 67663 Kaiserslautern, Germany

Abstract Quantum and quantum-inspired approaches offer emerging alternatives to conventional image segmentation methods. This paper provides a structured survey of existing approaches, covering quantum-inspired, tensor network-based, hybrid quantum-classical, and fully quantum methods. We compare these approaches and identify current challenges related to scalability, quantum hardware requirements, and reproducibility. The analysis highlights the predominance of quantum-inspired methods and medical image segmentation, while fully quantum approaches remain limited by current hardware constraints.

Keywords Image segmentation, quantum computing, quantum-inspired computing, quantum image processing, tensor networks

1 Introduction

Image segmentation is a foundational task in computer vision that partitions images into meaningful regions and supports applications ranging from medical imaging and autonomous driving to remote sensing,

while enabling downstream tasks such as object detection and scene understanding. Although modern deep learning methods, including convolutional and transformer-based architectures, achieve strong performance, they often require large annotated datasets, may suffer from limited interpretability and generalization, and incur substantial computational costs [1,2]. These limitations motivate the exploration of alternative computational paradigms, including quantum and quantum-inspired approaches that leverage quantum mechanical concepts such as superposition and entanglement.

This paper provides a compact overview of quantum image segmentation methods and organizes the literature into four categories: quantum-inspired, tensor network-based, hybrid quantum-classical, and fully quantum methods. The surveyed approaches are analyzed with respect to their underlying computational methods, image representations, application domains, and datasets, providing a structured comparison of the current quantum image segmentation landscape.

2 Background

Quantum Computing Basics Classical computation operates on bits, whereas quantum computation uses qubits whose states are described in a complex Hilbert space. A qubit can exist in a superposition of the computational basis states $|0\rangle$ and $|1\rangle$, described by complex probability amplitudes whose squared magnitudes sum to one. Multiple qubits form a joint quantum state whose dimension grows exponentially with the number of qubits, and measurement converts the state into classical outcomes according to these squared amplitudes. Quantum algorithms manipulate these states through sequences of quantum gates followed by measurement.

Quantum Image Representations Quantum image representations encode classical image data into quantum states. Among the most established approaches is the Novel Enhanced Quantum Representation (NEQR) [3], which explicitly encodes pixel intensities and spatial information using quantum registers. Several quantum and hybrid approaches also use other feature representations or encodings, such as angle, amplitude, intensity, and graph-based representations. The

choice of representation affects the required qubits, state-preparation procedure, circuit complexity, and measurement requirements.

Tensor Networks Tensor networks provide a structured representation of high-dimensional quantum states by decomposing them into interconnected lower-dimensional tensors. Common structures include Matrix Product States (MPS), Projected Entangled Pair States (PEPS) and Multi-scale Entanglement Renormalization Ansatz tensor network (MERA). The bond dimension, controlling tensor connectivity, determines the number of parameters and the resulting trade-off between representation accuracy and computational cost. Tensor networks can therefore provide compact representations of high-dimensional data and enable classical processing of structures originating from quantum-state representations [4–6].

3 Quantum Image Segmentation Methods

3.1 Methodology and Taxonomy

A structured literature search was conducted for quantum and tensor network-based image segmentation methods published between January 2019 and July 2026, primarily via Lens.org and complemented by Google Scholar, using the queries "Quantum AND Image AND Segmentation" and "Tensor AND Network AND Image AND Segmentation". AI-based tools served only as supplementary aids for literature discovery and citation tracking. Studies were included if image segmentation was a central task and the method incorporated quantum-mechanical principles, quantum computing, quantum-inspired methods, or tensor networks. No restriction was imposed on publication type; preprints were retained given the field's youth and heterogeneity, to capture recent and emerging developments. After title and abstract screening and full-text inspection where required, 36 publications were selected, comprising 16 quantum-inspired, 4 tensor network-based, 9 hybrid quantum-classical, and 7 fully quantum approaches, forming the basis for the taxonomy shown in Fig. 1.

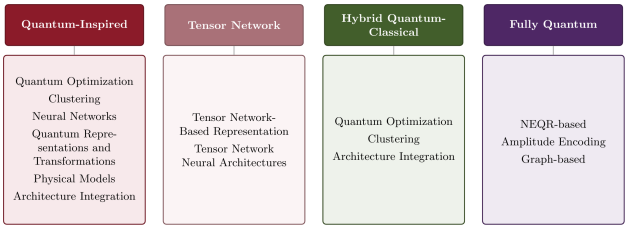


Figure 1: Four-category taxonomy of quantum and quantum-inspired image segmentation methods and their subcategories. The similar colors of quantum-inspired and tensor network methods indicate their relationship.

3.2 Quantum-Inspired Methods

The majority of the surveyed literature falls into the quantum-inspired category. These methods operate on classical hardware while incorporating concepts from quantum mechanics or their underlying mathematical frameworks. They can be further divided into the subcategories shown in Fig. 1, which are summarized in Table 1.

Table 1: Overview of quantum-inspired image segmentation methods.

Group	Literature	Method
Quantum-inspired optimization	[7,8]	Quantum-behaved particle swarm optimization
	[9]	Marine predator algorithm (Schrödinger-derived probability)
	[10]	Firefly algorithm with superposition
	[11]	Quantum-inspired optimal foraging algorithm
	[12]	Entropy-based thresholding
	[13]	Differential evolution + quantum-behaved PSO
Clustering	[14]	Dynamic quantum clustering (Schrödinger-derived potential)
	[15]	Fuzzy c-means with quantum-inspired uncertainty modeling
Neural network	[16]	Self-supervised network with quantum neurons
	[17]	Quantum-inspired self-organizing network
Quantum representations and transformations	[18]	QUBO formulation
	[19]	Quantum phases + inverse QFT
Physical models	[20]	Quantum walk for superpixel graph
	[21]	Anderson localization
Integration into conventional architectures	[22]	U-Net with entanglement-inspired skips + superposition-inspired activations

3.3 Tensor Network Methods

A subclass of quantum-inspired approaches comprises tensor network methods, which represent high-dimensional data through structured tensor decompositions. These can be divided into tensor network-based representations, where tensor networks provide compact image representations, and tensor network neural architectures, where they are integrated directly into segmentation networks.

In the first group, Selvan et al. [23] use Matrix Product States (MPS) for patch-based medical image segmentation, approximating a high-dimensional pixel classifier with shared MPS weights. Similarly, Kopriva and Sitnik [24] propose a semantic segmentation method based on a Multi-scale Entanglement Renormalization Ansatz (MERA) tensor network for medical hyperspectral images, generating pseudo-labels before classifying the remaining pixels using a nearest subspace classifier.

In the second group, Konar et al. [25] propose 3D-QNet, a self-supervised tensor neural network for volumetric segmentation that uses quantum neurons and tensor decomposition for voxel-wise processing. Verma et al. [26] extend this approach to qutrits and tensor ring decomposition, increasing the representational capacity.

3.4 Hybrid Quantum-Classical Methods

Hybrid quantum-classical methods combine classical processing with quantum optimization or quantum neural components.

Pramanik et al. [27] propose a hybrid pipeline for crack detection and segmentation that combines classical preprocessing with quantum k-means and variational quantum classifiers for clustering and region-wise classification.

Several approaches use quantum optimization to solve classical segmentation formulations. Le et al. [28] formulate geometric-constrained segmentation as a graph partitioning problem and encode it as a QUBO, whose ground state is estimated using QAOA [36]. Presles et al. [29] similarly combine classical optimization with quantum annealing in an Expectation-Maximization framework for synthetic aperture radar image segmentation. Venkatesh et al. [30] propose Q-Seg, which formulates pixel-wise segmentation as a graph-cut QUBO and solves it

Table 2: Overview of hybrid quantum-classical image segmentation approaches, indicating which parts are processed classically and which use quantum resources.

Group	Literature	Classical Part	Quantum Part	Computation
Clustering	[27]	Dimensionality reduction	Segmentation	AWS Braket Local Simulator
Quantum optimization	[28]	QUBO formulation	Segmentation	Simulation
	[29]	QUBO formulation	Segmentation	D-Wave
	[30]	QUBO formulation	Segmentation	D-Wave
	[31]	QUBO formulation	Segmentation	D-Wave
Architecture integration	[32]	Processing, Segmentation	Feature representation improvement	Simulation
	[33]	Processing, Segmentation	Global feature integration	PennyLane simulation
	[34]	Processing, Segmentation	Feature representation improvement	Simulation
	[35]	Processing, Segmentation	Improve generalization and robustness	Simulation

on a D-Wave Advantage quantum annealer. Jun and Lee [31] also use a D-Wave annealer for CT segmentation, formulating the reconstruction problem as a QUBO while reducing the required logical qubits using known material attenuation coefficients.

A second group integrates quantum components into neural segmentation architectures. Xu et al. [32] introduce QuantumUNet and QuantumHoverNet, which incorporate lightweight quantum layers and evaluate them using idealized quantum simulations. Lim et al. [33] replace the U-Net bottleneck with an eight-qubit variational circuit, while Hossain et al. [34] integrate a shallow parameterized quantum circuit into the bottleneck to enhance compressed encoder features. Finally, Jaincy and Pattabiram [35] propose Q-SpinalNet, combining a U-Net-style encoder–decoder with transformer attention, residual connections, and a simulated quantum neural network.

Table 2 additionally shows which parts of the segmentation pipeline are performed classically and which are executed on quantum hardware or simulators.

3.5 Fully Quantum Methods

The fourth category consists of fully quantum algorithms designed for execution on quantum systems, although some studies are evaluated

using simulators due to current hardware limitations. Table 3 summarizes the image encoding strategies used and reports the quantum hardware or simulator employed in their experiments.

Table 3: Overview of fully quantum image segmentation methods and the quantum hardware or simulator used for the experiments.

Group	Literature	Method	Computation
NEQR	[37]	Grayscale morphological operations, quantum binarization	IBM-Q
	[38]	Background difference	IBM-Q
	[39]	Avoid quantum RESET, ancilla qubits	Qiskit simulation
	[40]	Dual-threshold quantum comparator	IBM-Q
Amplitude encoding	[41]	SWAP-test measurements of reconstructed quantum states	PennyLane simulation
Graph	[42]	QAOA	Conceptual
	[43]	Parametric Gate, Ancilla Basis, and Adaptive Cost Encoding	QASM simulator

4 Comparison

Three comparison categories are considered in this survey: application field, image encoding and representation, and datasets. Regarding the application fields, most contributions focus on medical images (16), with all tensor network methods belonging to the medical domain [23–26], and general images (13); the substantial number of medical studies highlights the continued need for effective segmentation methods where annotated data are often scarce. Table 4 shows the belonging of the literature to all domains extracted from the literature.

Table 4: Application domains of quantum and quantum-inspired image segmentation approaches.

Application Domain	Literature
Medical	[8, 10–12, 14–16, 18, 22–26, 31, 32, 35]
General	[9, 13, 17, 19, 20, 28, 33, 37–40, 42, 43]
Anomaly Detection	[21, 27, 41]
Remote Sensing, Satellite	[29, 30, 34]
Sonar Images	[7]

Regarding image encoding and representation, most quantum-inspired approaches rely on intensity- or histogram-based mappings,

avoiding the need for explicit quantum state preparation, including phase-encoded pixel intensities [16, 19], graph-based representations [20], and a two-dimensional entropy-based representation [13]. For tensor networks, methods range from tensor ring decompositions and angle-based encodings [26] to patch-based representations [23], effectively leveraging the inherent structure of tensor networks.

Hybrid approaches generally avoid quantum-native encodings such as FRQI or NEQR, relying instead on classical feature extraction, such as CNN-based embeddings [27], U-Net feature maps [34], or intensity- and graph-based representations [28, 29]. Fully quantum methods, in contrast, rely heavily on dedicated quantum encodings such as NEQR [37–40].

Regarding datasets, some studies rely on standard datasets: INbreast [18], the Cancer Genome Atlas [14], and BraTS [25, 26] for medical imaging; MVTec AD [41] and the Kaggle surface crack dataset [27] for anomaly detection; and Flowers and PASCAL VOC 2012 [19] for general natural and synthetic images.

5 Conclusion

In conclusion, the review of the literature reveals several gaps that currently limit the development of quantum image segmentation.

A first gap concerns benchmarking and standardized datasets. The literature uses highly heterogeneous datasets, with many studies providing only general descriptions of their data, making direct comparison and reproducibility difficult.

A second gap concerns application coverage. General and medical images dominate the literature, while domains such as remote sensing and anomaly detection remain underexplored and fields such as robotics have not yet been investigated, despite their potential to benefit from reduced supervision requirements.

A third gap relates to image representation. While quantum-inspired and hybrid approaches can operate on classical, compressed, or graph-based representations, fully quantum methods often require dedicated quantum encodings, whose state-preparation and hardware requirements currently limit scalability to large-scale, realistic images.

Addressing these gaps through standardized benchmarks, broader

applications, scalable quantum representations, further exploration of tensor networks, and consistent classical baselines will be essential for advancing quantum image segmentation.

References

1. N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical image analysis*, vol. 63, 2020.
2. C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimedia Systems*, vol. 27, no. 4, pp. 599–613, 2021.
3. Y. Zhang, K. Lu, Y. Gao, and M. Wang, "Neqr: a novel enhanced quantum representation of digital images," *Quantum information processing*, vol. 12, pp. 2833–2860, 2013.
4. J. Biamonte, "Lectures on quantum tensor networks," *arXiv preprint arXiv:1912.10049*, 2019.
5. R. Orús, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," *Annals of Physics*, vol. 349, pp. 117–158, 2014.
6. U. Schollwoeck, "The density-matrix renormalization group in the age of matrix product states," *Annals of Physics*, vol. 326, no. 1, pp. 96–192, 2011.
7. Y. Guo, L. Wei, and X. Xu, "A sonar image segmentation algorithm based on quantum-inspired particle swarm optimization and fuzzy clustering," *Neural Computing and Applications*, vol. 32, no. 22, pp. 16 775–16 782, 2020.
8. S. A. Rather, A. Kandwal, M. K. Pandit, and P. P. Roy, "Quantum-behaved particle swarm optimization for the segmentation of kidney stone CT images," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, pp. 1–5.
9. M. Abd Elaziz, D. Mohammadi, D. Oliva, and K. Salimifard, "Quantum marine predators algorithm for addressing multilevel image segmentation," *Applied Soft Computing*, vol. 110, 2021.
10. A. Choudhury, S. Samanta, S. Pratihar, and O. Bandyopadhyay, "Multilevel segmentation of hippocampus images using global steered quantum inspired firefly algorithm," *Applied Intelligence*, vol. 52, no. 7, pp. 7339–7372, 2022.

11. G. I. Sayed, "A novel multilevel thresholding algorithm based on quantum computing for abdominal ct liver images," *Evolutionary Intelligence*, vol. 16, no. 2, pp. 439–483, 2023.
12. A. Tariq Jamal, S. Abdel-Khalek, and A. Ben Ishak, "Multilevel segmentation of medical images in the framework of quantum and classical techniques," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13 167–13 180, 2023.
13. A. A. Bahaddad, S. Abdel-Khalek, S. Alkhalaf, H. M. AbdelSalam, A. B. Ishak, and M. Aripov, "Elevating image segmentation with multilevel two-dimensional quantum representation," *Plos one*, vol. 20, no. 9, 2025.
14. J. Sanchez and M. M. Landrove, "A dynamic quantum clustering approach to brain tumor segmentation," *arXiv preprint arXiv:2107.07698*, 2021.
15. J. Rathee, P. Kaur, and A. Singh, "Advancing brain tumor detection in noisy environments: A quantum inspired MRI scan analysis using interval type-2 fuzzy sets," *Computational Intelligence*, vol. 42, no. 2, 2026.
16. D. Konar, S. Bhattacharyya, T. K. Gandhi, and B. K. Panigrahi, "A quantum-inspired self-supervised network model for automatic segmentation of brain mr images," *Applied soft computing*, vol. 93, 2020.
17. P. Pal, S. Bhattacharyya, and N. Agrawal, "Grayscale image segmentation with quantum-inspired multilayer self-organizing neural network architecture endorsed by context sensitive thresholding," in *Research Anthology on Advancements in Quantum Technology*. IGI Global, 2021, pp. 197–227.
18. L. Domingo and M. Chehimi, "Quantum-enhanced unsupervised image segmentation for medical images analysis," *arXiv preprint arXiv:2411.15086*, 2024.
19. T. A. Akinola, X. Li, R. Wilkins, P. Obiomon, and L. Qian, "Robust inverse quantum fourier transform inspired algorithm for unsupervised image segmentation," *IEEE Access*, 2024.
20. W.-M. Shi, F.-X. Xu, Y.-H. Zhou, and Y.-G. Yang, "A novel image segmentation algorithm based on continuous-time quantum walk using superpixels," *International Journal of Theoretical Physics*, vol. 63, no. 1, 2023.
21. A. Srinivasan, A. Geng, A. Macaluso, M. Kiefer-Emmanouilidis, and A. Moghiseh, "A comparative study of q-seg, quantum-inspired techniques, and u-net for crack image segmentation," *tm-Technisches Messen*, 2025.
22. N. Ahmad and M. M. S. Beg, "A quantum machine learning inspired U-Net (QisNet) for retinal image segmentation," in *2026 6th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, feb 2026, pp. 1–5.

23. R. Selvan, E. B. Dam, S. A. Flensborg, and J. Petersen, "Patch-based medical image segmentation using matrix product state tensor networks," *Machine Learning for Biomedical Imaging*, vol. 1, pp. 1–24, 2022.

24. I. Kopriva and D. Sitnik, "MERA tensor network extracted pseudo-labels and grassmann manifolds for unsupervised intraoperative semantic segmentation of hyperspectral images of adenocarcinoma of the colon in the liver," in *Medical Imaging 2026: Digital and Computational Pathology*, vol. 13932. SPIE, 2026, pp. 231–243.

25. D. Konar, S. Bhattacharyya, T. K. Gandhi, B. K. Panigrahi, and R. Jiang, "3-d quantum-inspired self-supervised tensor network for volumetric segmentation of medical images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 312–10 325, 2024.

26. P. Verma, H. Kumar, D. K. Shukla, S. Satpathy, D. M. Alsekait, O. I. Khalaf, A. Alzoubi, B. S. Alqadi, D. S. AbdElminaam, A. Kushwaha *et al.*, "V3dqutrit a volumetric medical image segmentation based on 3d qutrit optimized modified tensor ring model," *Scientific Reports*, vol. 15, no. 1, pp. 1–12, 2025.

27. S. Pramanik, M. G. Chandra, C. V. Sridhar, A. Kulkarni, P. Sahoo, C. D. V. Vishwa, H. Sharma, V. Navelkar, S. Poojary, P. Shah *et al.*, "A quantum-classical hybrid method for image classification and segmentation," in *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*. IEEE, 2022, pp. 450–455.

28. N. H. Le, M. Sonka, and F. Toor, "A quantum optimization method for geometric constrained image segmentation," *arXiv preprint arXiv:2310.20154*, 2023.

29. T. Presles, C. Enderli, G. Burel, and E. H. Baghious, "Synthetic aperture radar image segmentation with quantum annealing," *IET Radar, Sonar & Navigation*, vol. 18, no. 5, pp. 812–824, 2024.

30. S. M. Venkatesh, A. Macaluso, M. Nuske, M. Klusch, and A. Dengel, "Q-seg: Quantum annealing-based unsupervised image segmentation," *IEEE Computer Graphics and Applications*, 2024.

31. K. Jun and H. Lee, "Quantum optimization algorithms for CT image segmentation from x-ray data," *Scientific Reports*, vol. 15, no. 1, 2025.

32. H. Xu, P. Xiao, D. Li, S. Qiu, H. Zhong, and Q. Zhou, "Quantum-enhanced deep learning for robust nuclei segmentation in histopathological images," in *2025 4th International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, 2025, pp. 219–222.

33. S.-C. Lim and J.-C. Kim, "Hybrid quantum-classical U-Net for image segmentation: Breaking the parameter scaling barrier," *IEEE Access*, vol. 14, pp. 33 934–33 947, 2026.
34. M. A. Hossain, A. V. Patel, I. Vanani, and B. Banerjee, "HQ-UNet: A hybrid quantum-classical u-net with a quantum bottleneck for remote sensing image segmentation," *arXiv preprint arXiv:2604.27206*, 2026.
35. M. Jaincy D E and V. Pattabiraman, "Quantum-SpinalNet: A hybrid deep learning approach for mammographic breast cancer detection," *Frontiers in Digital Health*, vol. 8, 2026.
36. E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," *arXiv preprint arXiv:1411.4028*, 2014.
37. W. Liu, L. Wang, and M. Cui, "Quantum image segmentation based on grayscale morphology," *IEEE Transactions on Quantum Engineering*, vol. 3, pp. 1–12, 2022.
38. L. Wang, W. Liu, and Z. Deng, "A quantum segmentation algorithm based on background-difference method for neqr image." *Quantum Inf. Comput.*, vol. 23, no. 15&16, pp. 1291–1309, 2023.
39. A. Prodan, V.-I. Manta, and O. Zvorișteanu, "Implementation of a two-threshold quantum image segmentation circuit in NEQR format," *Mathematics*, vol. 14, no. 9, 2026.
40. Y. Dong, R. Yan, D. Mou, and F. Li, "Improved dual-threshold quantum image segmentation algorithm and simulation," *Physica Scripta*, vol. 99, no. 7, 2024.
41. M. F. Madeira, A. Poggiali, and J. M. Lorenz, "Quantum patch-based autoencoder for anomaly segmentation," in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 1. IEEE, 2024, pp. 259–267.
42. L. Tse, P. Mountney, P. Klein, and S. Severini, "Graph cut segmentation methods revisited with a quantum algorithm," *arXiv preprint arXiv:1812.03050*, 2018.
43. S. M. Venkatesh, A. Macaluso, M. Nuske, M. Klusch, and A. Dengel, "Qubit-efficient variational quantum algorithms for image segmentation," in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 1. IEEE, 2024, pp. 450–456.

Detection of surface defects on free-form components

Daniel Sopauschke, Thomas Dunker, and Erik Trostmann

Fraunhofer Institute for Factory Operation and Automation IFF,
Sandtorstrasse 22, 39106 Magdeburg

Abstract This paper addresses the detection of surface defects, such as missing material, impact marks, dings, and dents, on cast and forged components. These defects typically deviate locally by several tenths of a millimeter from the expected geometry and exceed the measurement uncertainty of dense optical 3D scanning. We propose curvature-based features and non-rigid registration combined with distance computations to distinguish defects from permissible form deviations in point cloud data. Specialized data structures, intelligent reduction, and approximation techniques reduce computation time and support inline inspection.

Keywords Surface defects, inspection, point clouds, curvature features, non-rigid registration

1 Introduction

When inspecting cast and forged components, defects such as impact marks, dings, dents, and missing material must be distinguished from intended geometric features and permissible form deviations. Optical measurement systems capture surfaces as dense point clouds. The objective is to compute features from these point clouds that reliably differentiate defects from acceptable geometry.

[1] reviews surface-inspection methods based on 3D point clouds and groups them into profile-based, template-matching, local-feature, multimodal, and deep-learning approaches. The two approaches presented here fall within curvature- and distance-based local-feature methods.

First, we analyze local curvature values and compare them with expected values derived from a CAD model or a golden sample (section 2.1). This approach is particularly suitable for defects at edges and in other highly curved regions. Second, we introduce a non-rigid registration method that distinguishes permissible form deviations from actual defects (section 2.2).

Existing non-rigid registration implementations [2][1] and [3][2] were tested successfully on point clouds containing tens of thousands of points but exceeded practical memory limits for point clouds with approximately 100,000 points or more. Earlier work [4][3] and a learning-based method [5][4] were also considered.

2 Methods

We distinguish two approaches to defect detection: local curvature-based analysis and non-rigid ICP-based distance analysis. Curvature-based detection is better suited to smaller regions, such as straight edges or highly curved areas, whereas the registration approach can handle larger and more complex shapes.

2.1 Curvature-based analysis

Defects such as buckles and dents alter the local shape of a surface. They cannot always be detected reliably through distances to a CAD model or golden sample because permissible global deviations may mask the defect. Local curvature analysis offers an alternative: at a defect, the measured curvature typically deviates from the expected model value. Curvature in point clouds is estimated by fitting a mathematical structure within a local neighborhood. A simple method derives an unsigned curvature value from the eigenvalues of the local covariance matrix [6]. Other methods fit a local sphere to estimate ex-

[1] https://zikai1.github.io/pub/CluReg/index.html
[2] https://github.com/AIT-Assistive-Autonomous-Systems/3D_nonrigid_tricubic_pointcloud_registration
[3] https://github.com/shubhamag/non_rigid_icp, https://github.com/golkir/optimal-step-nonrigid-icp
[4] https://huangjh-pub.github.io/publication/synorim/

trinsic curvature or a second-order surface to obtain the two principal curvatures.

For curvature-based defect detection, the surface is divided into regions. A key distinction is between simply curved surfaces with Gaussian curvature zero, such as cylindrical or conical surfaces, and doubly curved regions, such as corners. This subdivision is prepared from the CAD model or a representative sample because the required curvature calculations for the whole model are computationally intensive. During inspection, regions are assigned using fast distance queries to the model and the category of the nearest neighbor or the majority category among several nearest neighbors.

In simply curved regions, defects can be detected by analyzing the principal curvatures: the second principal curvature should be close to zero, and significant deviations may indicate a defect. Thresholds must account for noise, sensor motion, and component motion. Because curvature magnitudes vary across a component, a single global threshold is generally unsuitable. Possible alternatives are a ratio between expected and measured curvature or a comparison with local statistics obtained from several defect-free reference parts. Points exceeding the threshold are clustered by spatial proximity and similarly aligned normal vectors to prevent opposite sides of thin components from being merged. Compact clusters are retained, whereas elongated or very thin clusters are treated as likely artifacts.

Computing principal curvatures is expensive because a second-order surface must be fitted for each point [7]. Spatial data structures such as kd-trees [8] and parallel execution reduce neighborhood-search costs but remain insufficient for large datasets and inline inspection. Approximate computation provides further acceleration [9]: curvatures are computed only for a subset of points and assigned to their local neighborhoods. Approximation ratios such as 1:4 or 1:8 preserve sufficient accuracy for most analyses while reducing computation time by roughly the same factor. This method uses a specialized kd-tree memory layout [10], allowing subsets to be derived directly from the tree structure. In doubly curved regions, extrinsic curvature obtained from sphere fitting can be used. Approximation is especially beneficial in these regions because sphere fitting is even more expensive.

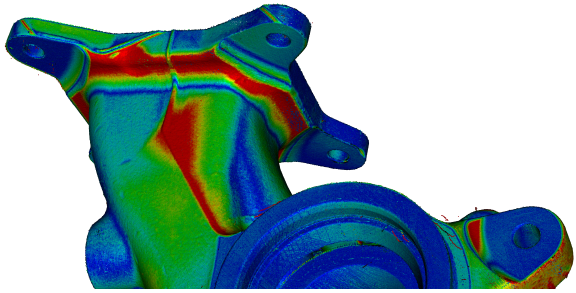


Figure 1: Differences between the scanned component and the CAD specification. The missing-material defect in the upper left is of the same order of magnitude as the permissible form deviation (Source: Fraunhofer IFF).

2.2 Non-rigid ICP-based distance analysis

A central challenge is distinguishing actual defects from permissible form deviations. Simple distances to the CAD specification are often insufficient. Registration spatially aligns datasets before further analysis. The standard Iterative Closest Point (ICP) algorithm [11] iteratively estimates rotations and translations from nearest-neighbor correspondences. This rigid alignment cannot compensate for permissible shape variation. Non-rigid registration instead deforms the scan so that large-scale permissible deviations are compensated while small defect regions are preserved. A subsequent distance calculation then emphasizes the actual defects.

The deformation follows an idea similar to Laplacian mesh smoothing [12]. For each scan point the nearest neighbor in the reference model is determined. Compatible normal vectors may be used to stabilize the correspondences search [13]. Each correspondence yields a displacement vector. Neighboring scan points are constrained to move coherently, enabling large-scale adjustment while preserving smaller surface details. This constraint is formulated as a large sparse linear system:

$$\begin{aligned} (\mathbf{W} + \lambda(\mathbf{D} - \mathbf{A})) \cdot \vec{\mathbf{U}} &= \mathbf{W} \cdot \vec{\mathbf{B}} \\ \mathbf{X}_{n+1} &= \mathbf{X}_n + \vec{\mathbf{U}} \end{aligned} \tag{1}$$

Here, $\mathbf{D} - \mathbf{A}$ is the Laplacian graph matrix; $\mathbf{A}$ is the sparse weighted adjacency matrix; $\mathbf{D}$ is the diagonal neighborhood-degree matrix; $\mathbf{W}$ adjusts the influence of individual displacement vectors; λ controls stiffness; $\mathbf{B}$ contains the initial displacements; and $\mathbf{U}$ contains the adjusted displacements applied to point cloud $\mathbf{X}_n$.

The neighborhood graph is computed once at the start. Neighborhoods are obtained either by radius search or k-nearest-neighbor search. Radius search is often easier to parameterize from component dimensions. Normal vector compatibility can prevent points on adjoining faces at sharp edges from being coupled too strongly. The search, sparse-system solution, and point-cloud update are repeated until convergence or a maximum iteration count is reached. Larger neighborhoods produce a stiffer deformation and better preserve small defects, but increase runtime.

For inline inspection, the original method is too slow: deforming a scan with approximately 500,000 points requires several minutes. The scan is therefore thinned homogeneously, deformation is computed on the reduced cloud, and final positions for the original points are reconstructed by inverse-distance weighting.

3 Results

The methods were evaluated on two datasets: a crankshaft with a small impact mark at an edge and a turbocharger housing with missing material. All tests were performed on a Windows 11 workstation equipped with an Intel Core i9-9900K processor and 64 GB RAM.

3.1 Curvature-based analysis

Figure 2 compares scan-to-reference distances with expected and measured second principal curvatures. The impact mark at the edge is masked by permissible deviations and even produces smaller distances than its surroundings. A purely distance-based method therefore fails. By contrast, the curvature maps reveal the defect clearly. Approximation ratios of 1:4 and 1:8 cause only minor differences without affecting detection, whereas 1:64 introduces excessive artifacts. For 3.8 million points and a curvature radius of 1.5, the 1:8 approximation reduced

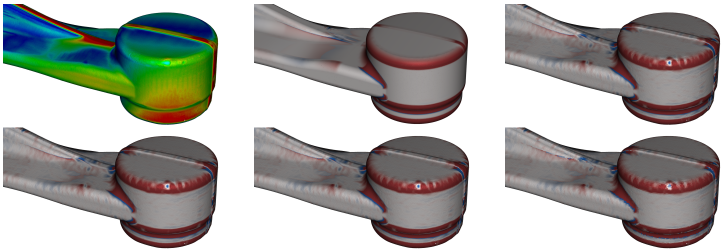


Figure 2: Top row: Distances between scan and reference; expected second principal curvatures; measured second principal curvatures at full resolution. Bottom row: Curvature estimates using 1:4, 1:8, and 1:64 approximations (Source: Fraunhofer IFF).



Figure 3: Residual distances after non-rigid registration. The missing-material defect is clearly visible against a background of compensated deviations (Source: Fraunhofer IFF).

runtime from approximately 60 s to 7 s. Inline operation is achieved by applying curvature analysis only to predefined relevant regions and excluding flat or otherwise irrelevant areas.

3.2 Non-rigid ICP-based distance analysis

Figure 3 shows the results of the non-rigid ICP method applied to the turbocharger from Figure 1. By allowing global, smooth deformation, permissible form deviations are compensated, leaving the defect location as the dominant deviation in the residual distance map.

A critical parameter is the number of iterations. Unlike rigid ICP,

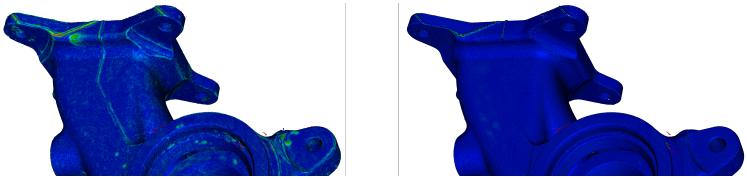


Figure 4: Adverse effects of misconfigured non-rigid registration: (Left) 20 iterations cause excessive smoothing; (right) too-small local neighborhood leads to over-alignment and loss of defect contrast (Source: Fraunhofer IFF).

where convergence typically requires dozens of iterations, our experiments show that five iterations suffice for non-rigid ICP. Excessive iterations (> 10) over-smooth the deformation, causing defects to blur until they become indistinguishable from noise. This is illustrated on the left of Figure 4, where 20 iterations reduce the defect's contrast without eliminating it.

The size of the local neighborhood used to build the Laplacian graph is equally important. An overly small neighborhood yields insufficient regularization, allowing the scan to overfit the reference: even small-scale noise may be compensated, thereby erasing defect signatures (right side of Figure 4). Larger neighborhoods enforce smoother deformations, better preserving defect features, albeit at higher computational cost. No universally optimal value exists, it must be tuned per component geometry. In general, a larger neighborhood is preferable.

To maintain inline capability, the original point cloud is thinned before registration and points are interpolated back afterward (Section 2.2). For the turbocharger (8.5 M points), computation time drops from 6 min 29 s to 2 s 761 ms, with no statistically significant loss in detection performance. Tables 1 and 2 quantify this acceleration. The additional effort for thinning and interpolation is more than compensated by the substantially reduced time of the remaining steps.

4 Summary

Detecting surface defects on free-form components is challenging because true defects must be separated from permissible shape variation. Two point-cloud methods were presented: curvature-based local

Table 1: Computation time breakdown for non-rigid ICP on the turbocharger point cloud (8.5 M points), with and without thinning. All times are wall-clock on an Intel Core i9-9900K.

Phase	**Full**	**Thinned**
Target tree creation	175ms	175ms
Source thinning	-	1s 32ms
Source tree creation	704ms	13ms
Laplacian graph matrix	6s 425ms	74ms
Flexible registration	6m 15s 118ms	581ms
Interpolation	-	886ms
Total	**6m 29s 726ms**	**2s 761ms**

Table 2: Detailed timings of the flexible registration phase for full and thinned version of the non-rigid ICP.

Flexible registration	**Full**	**Thinned**
Correspondences search	1s 438ms	29ms
Displacement solving	1m 18s 234ms	100ms
Displacement update	23ms	0.5ms
Residuals computation	18ms	0.25ms

analysis and non-rigid ICP-based distance analysis. In combination with suitable classification procedures, both methods localize defects. Particular emphasis was placed on computational acceleration through approximation, region-of-interest selection, point-cloud thinning, and interpolation, because short processing times are essential for inline quality inspection.

Acknowledgements

This research was sponsored by the Federal Ministry of Economic Affairs and Climate Action, grant number KK5195104RG4.

Supported by:

Federal Ministry for Economic Affairs and Climate Action

ZIM Zentrales Innovationsprogramm Mittelstand

on the basis of a decision by the German Bundestag

References

1. L. Huo, Y. Liu, Y. Yang, Z. Zhuang, and M. Sun, "Review: Research on product surface quality inspection technology based on 3d point cloud," *Advances in Mechanical Engineering*, vol. 15, no. 3, 2023.
2. M. Zhao, J. Jiang, L. Ma, S. Xin, G. Meng, and D.-M. Yan, "Correspondence-free non-rigid point set registration using unsupervised clustering analysis," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 21 199–21 208.
3. P. Glira, C. Weidinger, J. Otepka-Schremmer, C. Ressl, N. Pfeifer, and M. Haberler-Weber, "Nonrigid point cloud registration using piecewise tricubic polynomials as transformation model," *Remote Sensing*, vol. 15, no. 22, p. 5348, 2023.
4. B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.
5. J. Huang, T. Birdal, Z. Gojcic, L. J. Guibas, and S.-M. Hu, "Multiway non-rigid point cloud registration via learned functional map synchronization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2038–2053, 2023.
6. Gumhold, Stefan, Xinlong Wang, and Rob S. MacLeod, "Feature extraction from point clouds," *IMR*, pp. 293–305, 2001.
7. F. Khameneifar and H. Ghorbani, "On the curvature estimation for noisy point cloud data via local quadric surface fitting," *Computer-Aided Design and Applications*, vol. 16, no. 1, pp. 140–149, 2018.
8. J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.
9. S. Sauer, D. Sopauschke, T. Dunker, D. Berndt, and M. Heizmann, "5.3 Automated Registration of Large Assemblies with Robot Guided 3D Sensors and CAD-based Planning," *Proceedings - ettc2022*, pp. 129–136, 2022.
10. D. Sopauschke, C. Teutsch, E. Trostmann, and D. Berndt, "A parallel memory efficient outlier detection algorithm for large unstructured point clouds," in *Optical Measurement Systems for Industrial Inspection XII*. SPIE, 2021, p. 68.
11. P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

12. A. Nealen, T. Igarashi, O. Sorkine, and M. Alexa, "Laplacian mesh optimization," in *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, ser. ACM Conferences. New York, NY: ACM, 2006, pp. 381–389.

13. S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*. IEEE Comput. Soc, 2001, pp. 145–152.

STEP-ID: Computer vision-based part identification for post-production automation in additive manufacturing

Marc Gabriel Schneider, Lennart Metz, Ole Werger, Roman Seidel, and Anton Grabmaier

Fraunhofer Institute for Microelectronic Circuits and Systems IMS, Finkenstraße 61, 47057 Duisburg, Germany

Abstract Additive manufacturing contract producers fabricate small, frequently changing batches rather than long series. After printing, every part has to be identified and matched to its order before post-production can continue, yet customer-specific parts carry no markers that would allow this to be automated, so the step is still performed manually and becomes error-prone precisely when a batch contains geometrically similar parts from different orders. We present STEP-ID, a computer vision pipeline that automates it without requiring markers or retraining when the assortment changes. Given a photograph of parts on a surface, the system segments them, yielding the count needed to verify order completeness, and matches each part against a database of rendered STL models by embedding similarity. Two re-ranking stages refine this, one matching local features of the backbone, the other comparing real-world part size recovered from a calibrated camera; each stage improves on the result of the previous one. Where the batch composition is known, the assignment is resolved for all parts jointly by Hungarian matching instead of independently. On the ThingiPrint benchmark, local re-ranking raises Top-1 accuracy from 71.4 to 80.3 %. On a set of 50 self-printed industrial parts, most arranged in series whose members differ only in size, the full pipeline raises Top-1 accuracy from 29.1 to 73.9 %. Hungarian assignment further raises accuracy to 91.6 % on ThingiPrint and 86.4 % on our own parts when the batch composition is known.

Keywords Additive manufacturing, computer vision, instance segmentation, part identification, embedding-based retrieval

1 Introduction

Additive manufacturing (AM) contract producers routinely fabricate small, frequently changing batches of parts rather than long series of a fixed product. After printing, individual parts have to be identified, counted, and matched to the correct order, before being routed to the appropriate downstream step, such as support removal, surface finishing, or heat treatment. Because prototypes and parts are produced to customer-specific designs, they typically carry no physical markers (e. g. printed IDs or tags) that would allow automated identification, and adding such markers is often impractical or undesirable in a contract-manufacturing setting. In practice, this identification and matching step is therefore still often performed manually, which is time-consuming and error-prone, particularly when a print batch contains geometrically similar parts belonging to different customer orders.

Prior work showed that computer vision offers a promising way to automate this step by identifying a printed part directly from a photograph and matching it to its corresponding 3D model (STL file), without requiring physical markers, following two broad strategies. The first trains a classifier on the target assortment of parts [1], which is accurate but must be retrained whenever the assortment changes, i. e. for practically every new print job in contract manufacturing. The second trains a general vision encoder that produces embeddings suitable for similarity search, so that new parts are identified zero-shot by adding them to the embedding database [2, 3]. We pursue this second direction, since it requires training only once.

We build on the work of Mathioulakis et al. [2], which showed that a contrastively fine-tuned DINOv2 backbone [4] can identify printed parts from renderings of their STL files. That approach matches on the global CLS-token embedding alone, leaving the local patch descriptors that the backbone computes anyway unused. Exploiting local rather than only global features is a common approach for fine-grained retrieval, and it is established practice in landmark recognition, where the same difficulty arises: candidates are re-ranked using local descriptors, most recently by re-ranking transformers such as RRT [5] and R2-Former [6]. Another approach, ELViS [7], has been shown to be very lightweight and to generalize well to new domains; we there-

fore adopt this method, using the patch tokens exactly as the backbone emits them.

Beyond this, we make two further contributions. First, we distinguish two ways of resolving the final assignment. Independent Top-1 matching decides for each part separately, whereas Hungarian matching exploits knowledge of what has to be matched: after a print job it is typically known exactly which n parts the batch contains and in what quantities, so matching the batch as a whole permits a globally optimal assignment. Second, we address geometrically near-identical parts, where the appearance embedding alone cannot resolve which size within a series a part belongs to. Using a calibrated camera, as is common in industrial settings, we estimate the real-world size of each segmented part and compare it against the STL model, which yields a further substantial gain in accuracy.

2 Materials and Methods

Figure 1 gives an overview of the STEP-ID pipeline: instance segmentation of the input photograph, which also yields the part count used to verify order completeness (stage 1), embedding-based retrieval of the Top-k candidates (stage 2), and two re-ranking stages, local by patch tokens (stage 3) and size-based by measured part dimensions (stage 4).

2.1 Instance Segmentation (Stage 1)

We train RF-DETR-Seg XLarge [8] entirely on synthetic images generated with FLUX.2 [9], which we annotate automatically with SAM3 [10], exploiting its zero-shot labelling capability. This yields our training data without any manual annotation. Generating rather than collecting training images is viable here because we operate in an industrial setting with few distracting objects and largely predictable conditions, so the generated scenes remain realistic for the deployment case. Figure 2 shows examples, covering the range from parts neatly arranged on a conveyor to densely cluttered workbenches. We generate 1,000 images in total, 800 for training and 200 for validation, and train for 30 epochs with a batch size of 16 at a learning rate of 10^{-4}.

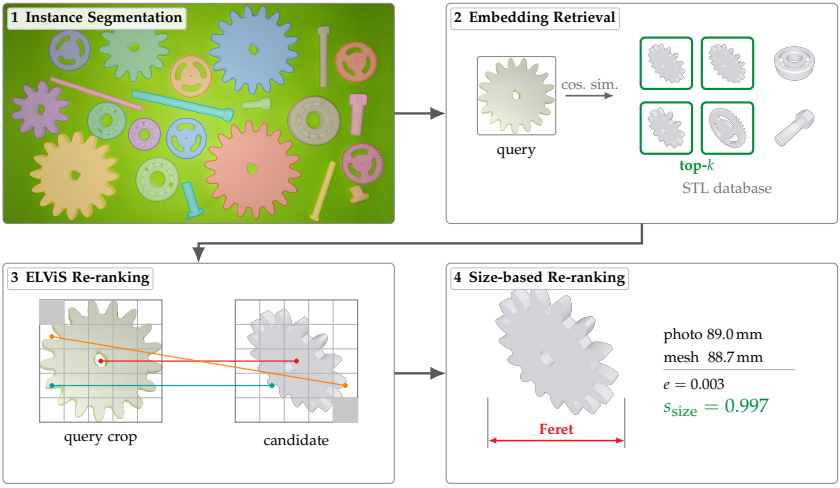


Figure 1: Overview of the STEP-ID pipeline, displaying the stages from input to match.



Figure 2: Generated training scenes for the segmentation model, spanning different surfaces, lighting conditions and packing densities.

2.2 Global Embedding-Based Retrieval (Stage 2)

As introduced above, we adopt this global retrieval stage from [2] and describe its training, rendering and inference setup below to the extent needed to understand it.

Training. We fine-tune the backbone using a contrastive learning approach, so that different views of the same object yield the same embedding, making it view-invariant, while different objects are pushed apart. We train with InfoNCE, which operates on pairs: we draw a ran-

dom anchor from an object's 24 renderings and pair it with exactly one positive, the adjacent azimuth view ($\pm 30^\circ$), with the remaining objects in the batch acting as negatives. Training uses all of Thingi10K [11] except the models appearing in our test set, giving 9,852 objects, each rendered in advance; individual renderings are then drawn from this pool during training. We train with 175 objects per batch for 40 epochs at a learning rate of 10^{-7}.

Rendering. The rendering scheme gives 24 renderings per model: two elevation bands at 30° and 60°, each sampled every 30° in azimuth. Renderings are produced at 448 px, matching the input resolution used at training and inference time. Each mesh is normalized to a maximum extent of one and placed at a fixed camera distance, so that all models fill the frame comparably regardless of their real-world size. For half of the renderings we additionally replace the plain white background with a procedural pattern in a random colour, in order to avoid the model relying on a uniform white background.

Inference. Since a part may be photographed in any orientation, no single rendering can represent a model on its own. We therefore embed all 24 renderings of a model, taking the CLS token of each as its descriptor, and average them into a prototype embedding that forms the model's database entry. This retains what the viewpoints agree on, the part's identity, and keeps the database at one vector per model. At inference, each segmented crop is embedded the same way, and the global score s_{global} is the cosine similarity between its CLS token and the prototype embedding of a candidate model, computed for the whole catalogue in a single matrix product.

Since the two re-ranking stages that follow are computationally more expensive, we re-rank only a shortlist of candidates rather than the whole catalogue. Rather than a fixed number, its size k^* is chosen dynamically from the top-K candidates $s_1 \geq \cdots \geq s_K$: we locate the sharpest normalized drop between consecutive scores,

$$k^* = \underset{k_{\min} \leq k < k_{\max}}{\arg\max} \frac{s_k - s_{k+1}}{\max(s_1 - s_K, \epsilon)}, \tag{1}$$

and keep the k^* highest-scoring candidates, i. e. everything up to and

including the model right before that drop, with a small constant $\epsilon > 0$ (we use $\epsilon \ll 1$) guarding against division by zero when the top-K scores are nearly identical. Only these candidates are passed on to the two re-ranking stages described next.

2.3 Local-Descriptor Re-Ranking with ELViS (Stage 3)

To recover the patch-level information discarded by the global CLS token, we re-rank the shortlist with ELViS [7], matching patch tokens between the query image and each candidate rendering directly in their similarity space. Not all patch tokens are equally informative for re-ranking: many correspond to background or to regions that are uninformative for distinguishing similar parts. The original ELViS formulation addresses this by training an additional detector network to select the most relevant patch tokens. We simplify this selection step: instead of training a separate detector, we inspect the CLS attention of the backbone's last transformer layer and select the top 25 % of patch tokens ranked by their attention weight to the CLS token. The underlying idea is that the patch tokens which most strongly shaped the global (CLS-token) decision are also the most relevant ones to re-examine explicitly for local correspondence during re-ranking. Only this reduced set of high-attention patch tokens is then used for the patch-to-patch matching step of ELViS, avoiding the cost of running a separate token-selection detector.

The local score s_{local} does not replace the global similarity of stage 2 but is combined with it into a single appearance score,

$$s_{\text{app}} = w_{\text{global}}\, s_{\text{global}} + (1 - w_{\text{global}})\, \text{sigmoid}(\gamma\, s_{\text{local}}), \tag{2}$$

where γ is a temperature that maps the unbounded matching score onto a comparable range.

2.4 Size-Based Re-Ranking (Stage 4)

From the convex hull of the segmented mask we compute both the maximum and the minimum Feret diameter, the largest and smallest extent of the shape across all orientations, and convert them to millimeters using the calibrated scale. We determine the same two quantities

for each candidate STL from the convex hull of its mesh vertices. Using only the maximum diameter is ambiguous, since a long, narrow part and a long, wide part can share it while differing in the other dimension; weighting both diameters equally when comparing photo and candidate avoids this confusion. We take the relative deviations $e_{\max}$ and $e_{\min}$ for each diameter and combine them into a single relative deviation $e = \frac{1}{2}(e_{\max} + e_{\min})$, mapped to a bounded agreement score $s_{\text{size}} = 1/(1+e)$. This treats size as a soft constraint rather than a cutoff, since a mismatch may also stem from a mis-segmented mask.

The cue is applied only to the shortlist, where it is fused with the appearance score of (2),

$$s_{\text{fused}} = w_{\text{size}}\, s_{\text{size}} + (1 - w_{\text{size}})\, s_{\text{app}}, \tag{3}$$

2.5 Matching

The stages above yield, for each query, a similarity score against every STL model. Matching each query independently means taking its highest-scoring model, which permits duplicates: two instances may be assigned the same model while another ordered model is left unassigned.

When the number of queries equals the number of ordered parts and each model is known to occur exactly once, we can instead optimize the assignment globally. Using the negative similarities as costs, we seek the one-to-one assignment of instances to models that minimizes the total cost, a linear assignment problem solved exactly by the Hungarian algorithm [12]. This lets an instance take its second-best candidate if doing so frees a model no other instance matches well, so that confident matches propagate through the ambiguous ones.

2.6 Benchmark Dataset

Our benchmark comprises 150 STL parts from two sources. The first is the public ThingiPrint benchmark [2], 100 parts selected from Thingi10K [11] with 10 real query photographs each, giving 1,000 queries against 100 reference models. It matches our photo-to-STL scenario, substantially enlarges the benchmark, and, containing objects that vary in appearance, measures how well the pipeline bridges the

render-to-photo domain gap. Lacking a calibrated scale reference, it serves to evaluate embedding-based retrieval only.

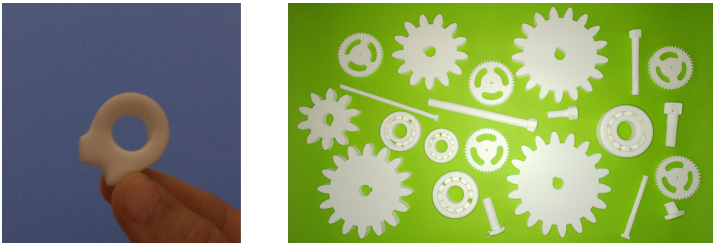


Figure 3: Query images from the two benchmark sources. **Left:** ThingiPrint, a single part held in hand. **Right:** Our own subset, an overhead shot of several parts at once, among them gears, bearings and screws.

The second source is 50 self-printed parts that we acquired to cover industrially relevant, highly similar geometries. Most belong to five series that are shape-identical within the series and differ only in scale: ball bearings from 33.0 to 55.0 mm in diameter, in steps of 5.5 mm; two visually distinct families of gears, printed to overlap in size by design, spanning 60.0 to 104.8 mm in diameter, in steps of about 5 mm, so that size alone cannot tell them apart; a third gear series packed far more tightly at 42.0 to 46.0 mm in diameter, in steps of 1 mm; and screws from M4 to M11, with lengths from 13.8 to 109.4 mm. The remainder are individual parts without a series partner. Within each series the members differ in size alone, so the appearance stages have to place the query in the right family and the size cue then has to resolve which member it is. We photograph the parts both individually and several at a time, segmenting the latter into individual crops before retrieval, giving 333 queries in total, on average 6.7 per model. The fixed camera is calibrated to a known pixel/mm scale using a reference plate with known dimensions.

3 Evaluation

Instance segmentation (stage 1). We first evaluate the segmentation stage on 209 of our query images, containing 373 instances of a single class (Table 1); Figure 3 (right) shows an example of such an image. Trained purely on synthetic generated images, RF-DETR-Seg XLarge transfers to the real acquisition setting essentially without loss: recall is 1.000 up to an IoU of 0.75, i. e. every instance is found. The drop under the stricter averaged criterion indicates that parts are reliably localized while mask boundaries are less exact at high IoU. Since the crops passed to retrieval are cut from these masks, boundary precision directly limits the size measurement of stage 4: comparing measured against ground-truth part size, we find a systematic bias of $+3.96\,\%$ of true part size, i. e. the pipeline tends to overestimate part size, with a standard deviation of $6.53\,\%$ of true part size across measurements. Consequently, stage 4 cannot reliably distinguish between parts whose true sizes differ by less than this margin.

Table 1: Instance segmentation on 209 query images with 373 instances of a single class.

Metric	Masks
mAP@0.75	0.987
mAP@[0.50:0.95]	0.833
AR@100, IoU 0.75	1.000
AR@100, IoU [0.50:0.95]	0.894

Retrieval accuracy across pipeline stages. Table 2 reports retrieval accuracy as the stages of the pipeline are added. The fusion weights w_{global}, w_{size} and the temperature γ are determined leave-one-class-out, so that no query is scored with parameters tuned on itself.

ThingiPrint shows clearly what fine-tuning and local re-ranking each contribute towards recognizing visually distinct parts more precisely. With 100 STL models, more than our own subset, it is also the more representative benchmark for how well ELViS works in general: on our own parts, discussed below, its effect is markedly weaker.

Figure 4 shows a case from ThingiPrint where ELViS corrects an incorrect global match by matching patch tokens between query and can-

Table 2: Retrieval accuracy in percent across pipeline stages, with re-ranking hyperparameters determined leave-one-class-out. Stage 4 requires a calibrated imaging setup and is therefore only available on our own subset.

Stage	Top-1	Top-3	Top-5	Top-10	MRR
ThingiPrint					
Stage 2, global, no fine-tuning	59.4	81.6	87.8	94.1	72.2
Stage 2, global, fine-tuned	71.4	90.6	95.2	98.4	81.9
+ Stage 3, local re-ranking (ELViS)	**80.3**	**94.9**	**96.6**	98.4	**87.7**
Ours					
Stage 2, global, no fine-tuning	24.9	42.6	56.5	74.2	39.6
Stage 2, global, fine-tuned	29.1	50.5	69.7	93.1	46.2
+ Stage 3, local re-ranking (ELViS)	25.5	52.9	72.7	90.4	45.4
+ Stage 4, size-based re-ranking	**73.9**	**94.9**	**97.6**	99.7	**84.9**
Size only	59.8	90.4	98.8	99.7	74.9

didate: stage 2 puts a model first that shares the query's silhouette but not its structure, whereas the patch-token correspondences of stage 3 concentrate on the distinctive geometry, here the two arms and the central body, and move the correct model to the top.

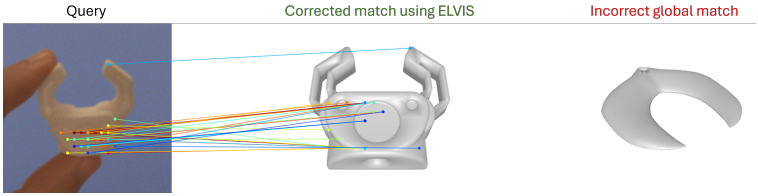


Figure 4: A global mismatch resolved by local re-ranking. **Left:** Query photograph and the model promoted by ELViS, with the established patch-token correspondences. **Right:** The model ranked first by the global embedding alone.

Our own subset behaves differently. Local re-ranking there lowers Top-1, from 29.1 to 25.5 %, while Top-3 and Top-5 still improve markedly. For a catalogue of visually near-identical parts differing only in size this is expected: ELViS cannot reliably tell which member of a series a query is and may promote the wrong one to first place, but it does make the correct model more likely to be among the top few, which is what the subsequent size comparison needs.

In a setting with a calibrated camera, knowing the real-world size of a part is a strong signal on its own, and clearly a better one than appearance for a catalogue of this kind: with a finite number of reference models, size alone already reaches 59.8 % Top-1, well above the 29.1 % of the best appearance-only stage. It cannot, however, distinguish between parts that share a size, which is exactly where the appearance-based pre-filtering of stages 2 and 3 helps, narrowing the shortlist down to parts of one type before the size comparison decides among them. This is why the combined stage 4 is the strongest by a wide margin, lifting Top-1 to 73.9 %.

Batch assignment. Table 3 evaluates the Hungarian assignment of Section 2.5 against independent Top-1 matching. Resolving the batch as a whole raises matching accuracy substantially on both subsets: since duplicate assignments are forbidden and every part is matched jointly rather than independently, optimizing the assignment globally recovers correct matches that independent Top-1 decisions miss.

Table 3: Best independent Top-1 accuracy versus Hungarian batch assignment, in percent, mean and standard deviation over 10 runs. Hungarian matching additionally assumes a known, complete batch, so the two columns are not directly comparable.

Subset	Top-1	Hungarian
ThingiPrint	80.3	**91.6** $\pm$ 2.6
Ours	73.9	**86.4** $\pm$ 5.1

4 Conclusion

We presented STEP-ID, a pipeline that identifies 3D-printed parts from a photograph and assigns them to a customer order without physical markers and without retraining when the assortment changes. It segments the parts, retrieves candidates from a database of rendered STL models, and refines the result in two re-ranking stages, one matching the local features of the backbone and one comparing the real-world part size recovered from a calibrated camera.

The evaluation shows that these two cues address different failure cases. On ThingiPrint, whose objects vary in appearance, local re-ranking is what carries the improvement, raising Top-1 accuracy from 71.4 to 80.3 %. On our own parts it lowers Top-1, while still improving Top-3 and Top-5, because a catalogue of size-graded series contains members that patch-token matching cannot separate. Size-based re-ranking then resolves precisely these cases and lifts Top-1 accuracy from 25.5 to 73.9 %, though it remains limited by segmentation mask precision. Neither cue is sufficient on its own: size alone reaches 59.8 % and appearance alone 29.1 %, since many parts share a size just as many share a shape, and their combination performs substantially better than either alone. Where a complete batch is known in advance, Hungarian assignment further improves accuracy on both subsets by resolving all parts jointly instead of independently, from 80.3 to 91.6 % on ThingiPrint and from 73.9 to 86.4 % on our own parts.

Acknowledgements

This research was completed as part of the Center for Applied Artificial Intelligence Duisburg (ZaKI.D). ZaKI.D is funded by the Federal Ministry for Economic Affairs and Energy (BMWi) and the Ministry of Economic Affairs, Industry, Climate Protection, and Energy of the State of North Rhine-Westphalia (MWIKE NRW) in the 5-location program (funding number 46SK0246A/B/C).

References

1. M. Meboldt *et al.*, "Recognition of additive manufacturing parts based on neural networks and synthetic training data: A generalized end-to-end workflow," *Applied Sciences*, vol. 13, no. 22, p. 12316, 2023.
2. D. Mathioulakis *et al.*, "Classifying novel 3d-printed objects without retraining," *arXiv preprint arXiv:2603.07465*, 2026.
3. Y. Liu, C. Graf, M. Spies, and M. Keuper, "Efficient fine-grained image retrieval with vision foundation models for industrial objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2026, pp. 3111–3120.

4. M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.
5. F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
6. S. Zhu *et al.*, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
7. P. Šuma, G. Kordopatis-Zilos, Y. Kalantidis, and G. Tolias, "Elvis: Efficient visual similarity from local descriptors that generalizes across domains," *ArXiv*, vol. abs/2603.28603, 2026.
8. Roboflow, "Rf-detr: Neural architecture search for real-time detection transformers," *arXiv preprint arXiv:2511.09554*, 2025.
9. Black Forest Labs, "Flux.2: Frontier visual intelligence," https://bfl.ai/blog/flux-2, 2025.
10. N. Carion *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.
11. Q. Zhou and A. Jacobson, "Thingi10k: A dataset of 10,000 3d-printing models," *arXiv preprint arXiv:1605.04797*, 2016.
12. H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

Zeitaufgelöste Brechungsindexentwicklung in DLP-gehärteten Photopolymeren für mikrooptische Anwendungen

Time-resolved refractive index formation in DLP-cured photopolymers for micro optical applications

Andreas Heinrich

Hochschule Aalen, Aalen School of Applied Photonics,
Anton Huber Straße 21, 73430 Aalen, Deutschland

Zusammenfassung Die präzise additive Fertigung mikrooptischer Komponenten, wie Gradientenindex (GRIN)-Optiken und diffraktiver optischer Elemente (DOEs), erfordert ein grundlegendes Verständnis der räumlich-zeitlichen Brechungsindexdynamik während der Photopolymerisation. Diese Studie präsentiert einen kombinierten experimentellen Ansatz zur quantitativen Analyse der Reaktionskinetik und der Entwicklung der optischen Eigenschaften eines acrylatbasierten Photopolymers während der Ein-Photonen-UV-Polymerisation. Durch die synchrone Messung der absoluten Brechungsindexänderung mittels Totalreflexion an einer Prismengrenzfläche und die Erfassung der Volumenphasengitterbildung durch Laserbeugung stellen wir eine direkte Korrelation zwischen Bestrahlungsstärke, Reaktionskinetik und optischer Funktionalität her. Die empirischen Ergebnisse zeigen eine S-förmige Brechungsindexentwicklung, die durch Sauerstoffinhibierung, Autoakzeleration und Vitrifizierung bestimmt wird. Darüber hinaus offenbart die zeitaufgelöste Beugungsanalyse komplexe Intensitätssubstrukturen, die von den diskreten Totzonen des Digital Light Processing (DLP)-Projektors herrühren. Ein auf Fourier-Optik basierendes Simulationsmodell reproduziert diese Beugungsmuster erfolgreich. Zeitliche Divergenzen zwischen der Simulation und den experimentellen Daten in der späten Aushärtungsphase bestätigen jedoch, dass optische Streuung und Radikaldiffusion eine verzögerte Polymerisation in nominell unbelichteten Bereichen

induzieren, was den Brechungsindexkontrast zunehmend verschlechtert.

Schlüsselwörter Photopolymerisation, Brechungsindexdynamik, Reaktionskinetik, Phasengitter

Abstract The precise additive manufacturing of micro-optical components, such as gradient-index (GRIN) optics and diffractive optical elements (DOEs), requires a fundamental understanding of spatiotemporal refractive index dynamics during photopolymerization. This study presents a combined experimental approach to quantitatively analyze the reaction kinetics and optical property evolution of an acrylate-based photopolymer during single-photon UV polymerization. By synchronously measuring the absolute refractive index change via total internal reflection at a prism interface and capturing the volume phase grating formation via laser diffraction, we establish a direct correlation between irradiance, reaction kinetics, and optical functionality. The empirical results demonstrate an S-shaped refractive index evolution governed by oxygen inhibition, autoacceleration, and vitrification. Furthermore, time-resolved diffraction analysis reveals complex intensity substructures originating from the discrete dead zones of the digital light processing (DLP) projector. A Fourier-optics-based simulation model successfully replicates these diffraction patterns. However, temporal divergences between the simulation and experimental data in the late curing phase confirm that optical scattering and radical diffusion induce delayed polymerization within nominally unexposed regions, progressively degrading the refractive index contrast.

Keywords Photopolymerization, refractive index dynamics, reaction kinetics, phase grating

1 Einleitung

Photopolymerisation ist ein grundlegender Mechanismus in der additiven Fertigung von optisch transparenten Mikroelementen [1, 2]. In Ein-Photonen-Stereolithographieverfahren, die Digital Light Processing (DLP)-Systeme nutzen, bestimmt die lokale UV-Dosis die

Monomer-zu-Polymer-Umwandlungsrate. Der Austausch von weitreichenden Van-der-Waals-Abständen gegen kurzreichweitige kovalente Bindungen induziert eine volumetrische Schrumpfung und einen proportionalen Anstieg der Materialdichte. Gemäß der Lorentz-Lorenz-Gleichung,

$$\frac{n^2-1}{n^2+2} = K \cdot \rho \tag{1}$$

wobei n der Brechungsindex, ρ die Dichte und K die molare Refraktivität ist, manifestiert sich diese Verdichtung physikalisch als lokale Erhöhung des Brechungsindex, zusätzlich zur Änderung der molaren Refraktivität [3].

In radikalischen Polymerisationssystemen ist die Erfassung dieser dynamischen Entwicklung aufgrund nichtlinearer kinetischer Phänomene eine Herausforderung. Im Harz gelöster molekularer Sauerstoff wirkt als starker Reaktionsinhibitor, indem er primäre Radikale abfängt und dadurch eine Induktionsperiode erzeugt. Nach Überschreiten dieses Schwellenwerts beschleunigt sich die Reaktionsrate typischerweise dramatisch (Trommsdorff-Norrish-Effekt), bevor sie stoppt, wenn das Material in einen glasartigen Zustand übergeht (Vitrifizierung).

Ziel dieser Studie ist es, diese kinetischen Phasen und ihre direkte Umsetzung in diffraktive optische Eigenschaften phänomenologisch zu charakterisieren und zu simulieren, um 3D-Druck-Belichtungsstrategien zu optimieren.

2 Material und Methoden

Die experimentellen Untersuchungen wurden mit dem kommerziell erhältlichen UV-härtbaren Harz PR48 (Autodesk) durchgeführt, das aus zwei Oligomeren (39,776 % Sartomer SR 494 LM, 39,776 % Allnex Ebecryl 8210), einem reaktiven Verdünner (19,888 % Rahn Genomer 1122), 0,4 % Esstech TPO+ als Typ-I-Photoinitiator und 0,16 % Mayzo OB+ als UV-Blocker formuliert ist.

Basierend auf der molaren Masse und Dichte beträgt der theoretische anfängliche Brechungsindex $n = 1,47$. Unter Anwendung von

Jacobs' Arbeitskurvenmodell [1] weist das Material eine Eindringtiefe von $D_p = 64\ \mu\text{m}$ und eine kritische Energie von $E_c = 17{,}3\ \text{mJ/cm}^2$ bei einer Bestrahlungsstärke von $20\ \text{mW/cm}^2$ auf.

Abbildung 1 zeigt den experimentellen Aufbau. Die zentrale Komponente ist ein Prisma, auf dessen Oberfläche das Polymer aufgetragen wurde. Ein Laserstrahl wird auf die Grenzfläche zwischen dem Polymer und dem Prisma ($n = 1{,}78$) fokussiert. Durch die Fokussierung treffen Lichtstrahlen aus einem bestimmten Winkelbereich auf die Probe. Abhängig von den Bedingungen für die Totalreflexion werden diese Lichtstrahlen entweder totalreflektiert oder aus dem Prisma ausgekoppelt. Durch Bestimmung des kritischen Winkels kann auf den Brechungsindex der Probe geschlossen werden. Dazu wird das totalreflektierte Licht von einer Kamera erfasst. Dadurch kann der Totalreflexionswinkel – und damit direkt der lokale Brechungsindex der Probe – aus der Position der Kante des Hell-Dunkel-Übergangs bestimmt werden.

Um Brechungsindexänderungen während der Polymerisation zu untersuchen, wird von unten durch das Prisma eine UV-Belichtungsmaske in das Polymer projiziert (Wintech4500 / DLP-Projektionssystem). Die Belichtungsmaske besteht aus abwechselnden Ein/Aus-Pixelreihen. Dadurch wird ein streifenförmiges UV-Muster in das Polymer projiziert, was zu einer streifenförmigen Aushärtung des Polymers führt. Konkret besteht eine „Ein-Reihe" aus zwei Pixelreihen (Pixelgröße: $40\ \mu\text{m} \times 40\ \mu\text{m}$) des DLP-Projektors. Aufgrund seiner inhärenten Totzone erzeugt ein einzelnes Pixel dabei keine homogene Ausleuchtung.

Zusätzlich wird ein zweiter Laserstrahl (Laser 2) auf das Polymer gerichtet und entsprechend dem sich im Laufe der Zeit entwickelnden Brechungsindexprofil gebeugt. Dies kann mit einer weiteren Kamera, die oberhalb des Prismas positioniert ist, aufgezeichnet und analysiert werden.

Um die Beugungsintensitäten mathematisch zu modellieren, wurde eine Fourier-Optik-Simulation in der Fraunhofer-Näherung erstellt [4]. Demnach gilt für das Feld in der Beobachtungsebene:

$$U_2(x,y) = \frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{k}{2z}(x^2+y^2)\right] \times \mathcal{F}\{U_1(x,y)\} \tag{2}$$

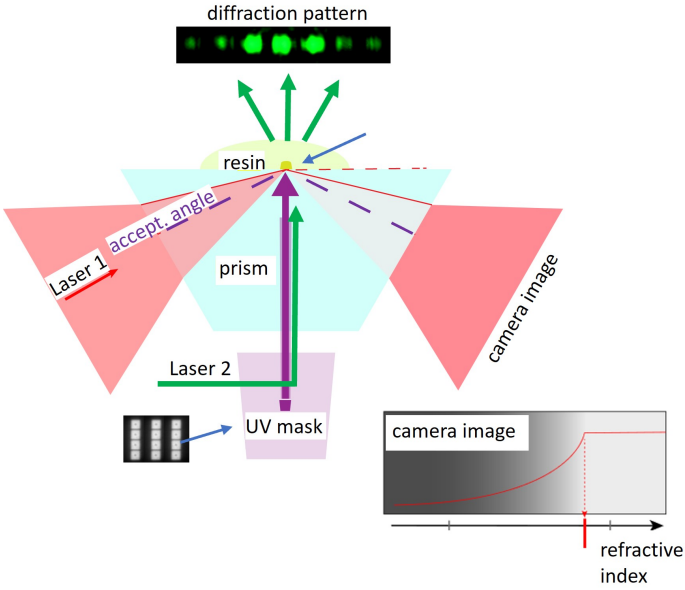


Abbildung 1: Schematische Darstellung des kombinierten Messaufbaus für Totalreflexion und Beugung.

wobei λ die Wellenlänge, z der Abstand von der Kamera zum Beugungsort und U_1 das Feld am Beugungsort ist. Entsprechend gilt für die Intensitätsverteilung in der Beobachtungsebene:

$$I_2(x,y) = \left(\frac{1}{\lambda z}\right)^2 \left(\mathcal{F}\{U_1(x,y)\}\right)^2 \tag{3}$$

und für die Beugungsebene ergibt sich:

$$U_1(x,y) = U_0 \cdot e^{j\phi(x,y)} \tag{4}$$

wobei die Phase gegeben ist durch:

$$\phi = \frac{2\pi}{\lambda} \Delta n(x,y,z,t) z(x,y,t) \tag{5}$$

Somit hängt die Intensität von der Brechungsindexdifferenz zwischen den belichteten und unbelichteten Bereichen und der Höhe des ausgehärteten Bereichs ab. Dabei resultiert die Intensität am Beobachtungspunkt aus der Überlagerung aller elektromagnetischen Wellen, die aus dem Volumen des Polymers kommen und am Beobachtungspunkt interferieren. Aufgrund der ortsabhängigen Polymerisation erfahren diese Wellen eine ortsabhängige Phasenverschiebung, da sich der Brechungsindex innerhalb des Volumens über die Zeit lokal ändert. Dementsprechend ist Δn eine Funktion der drei Raumkoordinaten und der Zeit.

Während des Übergangs vom flüssigen in den festen Zustand zeigt sich, dass die Brechungsindexänderung Δn lateral einen S-förmigen Verlauf aufweist. Zusätzlich ist die Entwicklung des Brechungsindexprofils in z-Richtung entscheidend für die Ausbildung der Intensitäten der Beugungsordnungen. Dies ist im verwendeten Versuchsaufbau nicht zugänglich, es kann jedoch davon ausgegangen werden, dass es ebenfalls einen S-förmigen Verlauf aufweist. Dabei muss jedoch berücksichtigt werden, dass gemäß dem Lambert-Beer-Gesetz die Intensität der UV-Projektion mit zunehmender Tiefe abnimmt, was zu einer reduzierten Aushärtung führt.

In der Simulation wird die Zeit in diskreten Schritten durchlaufen. Die für den jeweiligen Zeitschritt gültige Brechungsindexdifferenz ist dann entsprechend durch $\Delta n(z,t)$ gegeben und dient als erste Eingangsgröße zur Berechnung der zeitabhängigen Phase.

Da die Photopolymerisation aufgrund der maskenförmigen Belichtung auftritt, muss dies als Formfaktor abgebildet werden, um das Höhenprofil $z(x,y,t)$ (Aushärtungsfront) für die Phasenberechnung abzuleiten. Dazu wurde zunächst die Form ausgehärteter Pixel mittels Weißlichtinterferometrie erfasst. Dabei sind DLP-Projektor-inhärente Totzonen zwischen den einzelnen Pixeln und in der Mitte eines Pixels zu erkennen, die eine reduzierte Aushärtung des Polymers in diesen Bereichen verursachen. Für die zentrale, kreisförmige Totzone führt dies zu einer Höhenreduzierung von etwa 40 % im Vergleich zum Plateau.

Das Profil eines einzelnen ausgehärteten Pixels wurde dann in einen normierten Formfaktor übersetzt. Das Modell vernachlässigte die Tatsache, dass die Pixel an ihren Eckpunkten weniger Aushärtung verursachen als an den Mittelpunkten der Seitenkanten. Der Formfaktor für

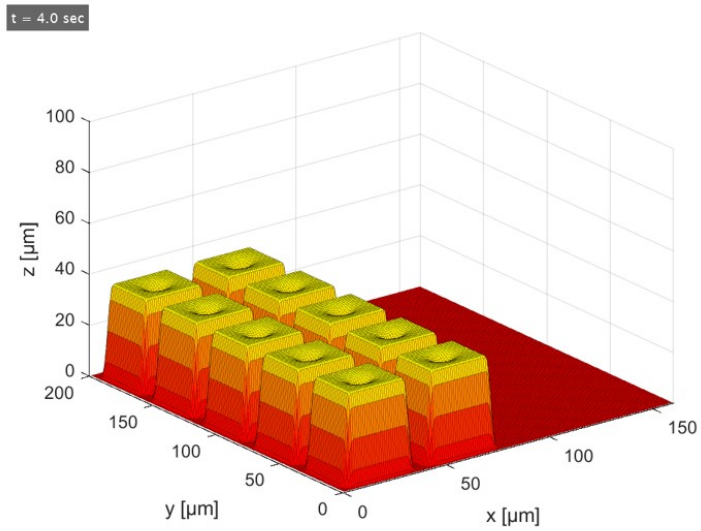


Abbildung 2: Ausschnitt des 3D-Formfaktors bzw. der Aushärtungsfront nach 4 s Aushärtung.

ein einzelnes Pixel ist somit von einer vollständigen Totzone umgeben. Die so für ein einzelnes Pixel erzeugte Form wurde repliziert, um ein entsprechendes Gitter zu erzeugen.

Schließlich wurde die höhennormierte 3D-Funktion erhalten. Dafür wird der Formfaktor in einen 3D-Formfaktor umgewandelt, der über die Simulationszeit zunimmt und die Bewegung der Aushärtungsfront in das Polymer beschreibt. Aufgrund der Kinetik bewegt sich die Aushärtungsfront jedoch während der Polymerisation nicht linear in das Polymer, was ebenfalls berücksichtigt werden muss. Abbildung 2 zeigt die modellierten Höhen des ausgehärteten Profils bei 4 s.

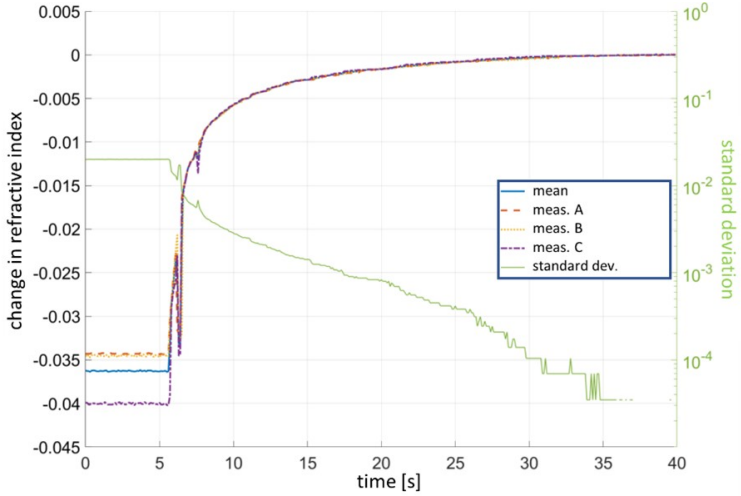


Abbildung 3: Zeitabhängige Entwicklung des Brechungsindex unter kontinuierlicher Bestrahlung bei 10 mW/cm^2.

3 Ergebnisse

3.1 Brechungsindex und Reaktionskinetik

Kontinuierliche UV-Bestrahlung (10 mW/cm^2) an der Prismengrenzfläche löst eine S-förmige Brechungsindexantwort aus (siehe Abbildung 3). Nach einem anfänglichen Sauerstoffinhibierungs-Plateau führt die Autoakzeleration zu einem steilen Indexanstieg, der letztendlich eine gesamte Brechungsindexverschiebung von $\Delta n \approx 0,04$ vor der Vitrifizierung ergibt. Kinetische Auswertungen zeigen dabei, dass die Sauerstoffinhibierungszeit exponentiell mit zunehmender einfallender Bestrahlungsstärke abnimmt, während die Propagationsrate eine direkte lineare Skalierung aufweist [4].

3.2 Entwicklung des Beugungsmusters

Abbildung 4 zeigt eine typische Entwicklung der Beugungsmuster. Der Laserspot wurde im Zentrum des UV-Maskenbildes positioniert (sie-

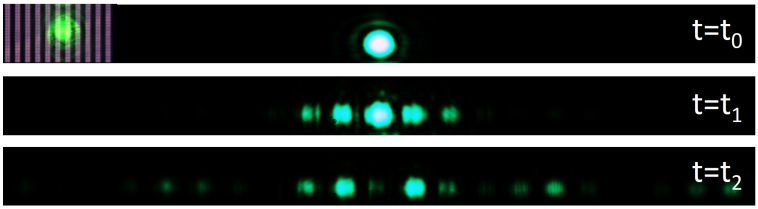


Abbildung 4: Zeitabhängige Entwicklung des Beugungsmusters unter kontinuierlicher Bestrahlung bei 10 mW/cm^2.

he ïnset"). Zum Startzeitpunkt t_0 liegt eine homogene Brechungsindexverteilung im gesamten flüssigen Polymer vor, sodass der Laser die Probe ohne Beugung durchdringt. Mit fortschreitender Zeit wird sichtbar, dass höhere Ordnungen auftreten und wieder verschwinden. Grundsätzlich weist jede Beugungsordnung eine Substruktur von 3 Maxima und 2 Minima auf, die jedoch aufgrund von Unschärfe nicht immer sichtbar ist.

Zusätzlich wird bei Verwendung höherer Belichtungszeiten der Kamera deutlich, dass das Phasengitter letztendlich in zwei Dimensionen beugt, was auf die 2D-Pixelstruktur der Projektorbeleuchtung zurückzuführen ist. Sobald die Probe vollständig ausgehärtet ist, friert das Beugungsmuster entsprechend ein. In Randbereichen des Gitters treten zudem reduzierte Kinetiken aufgrund von Temperaturgradienten und fehlenden Nachbarpixeln auf.

In Abbildung 5 ist der zeitliche Intensitätsverlauf der einzelnen Ordnungen dargestellt (15 Hz Kamera-Ausleserate). Zum Zeitpunkt $t = 0$ wurde die UV-Maskenbeleuchtung aktiviert. Nach etwa 1,2 Sekunden wird eine Änderung in der Beugungsstruktur sichtbar. Die Intensität der 0. Ordnung nimmt ab. Dementsprechend steigt die Intensität der 1. und 2. Beugungsordnung aufgrund der Energieumverteilung. Mit fortschreitender Zeit nimmt die Phasenverschiebung zwischen den unbelichteten und belichteten Bereichen weiter zu, was letztlich dazu führt, dass die Intensität der 0. Beugungsordnung ein Minimum durchläuft und sich wieder einem Maximum nähert. Gleichzeitig zeigen die Intensitäten in den anderen Beugungsordnungen das entgegengesetzte Verhalten.

Läge ein ideales Phasengitter vor, wie es z. B. durch die Raman-

Nath-Theorie beschrieben wird, würde sich das Verhalten für alle Beugungsordnungen oszillierend wiederholen. Nach etwa 3 Sekunden tritt jedoch eine Abweichung vom erwarteten Verhalten auf. Die zweite Beugungsordnung zeigt keinen symmetrischen Verlauf mehr. Ähnlich zeigen die dritte Beugungsordnung bei 5 Sekunden, die erste Beugungsordnung bei 7 Sekunden und die nullte Beugungsordnung bei 9 Sekunden einen asymmetrischen Verlauf.

3.3 Simulation von Phasengittern

Das angewendete numerische Modell basierend auf Fourier-Optik synthetisiert die oszillierenden Intensitätsverteilungen der Ordnungen erfolgreich (siehe gestrichelte Linie in Abbildung 5). Wie im Experiment wird zu Beginn der Aushärtung nur ein Signal der 0. Beugungsordnung beobachtet. Dieses Signal nimmt ab, während sich das Phasengitter entwickelt, wohingegen die Signale der 1. und 2. Beugungsordnung zunehmen.

4 Diskussion

Der exponentielle Abfall der Inhibierungsphase veranschaulicht einen diffusionskontrollierten Bereich, der bei niedrigen Intensitäten durch die Sauerstoffzufuhr aus der Atmosphäre in das Polymer dominiert wird. Bei höheren Photonenflüssen werden diese freien Sauerstoffmoleküle durch die photoinitiierten Radikale (TPO) vollständig abgebaut und die Photopolymerisation kann beginnen. Aufgrund der anfänglich hohen Mobilität der Monomere beschleunigt sich dieser Prozess, wodurch der Brechungsindex schnell ansteigt. Mit fortschreitender Polymerisation nimmt die Mobilität der Monomere und Polymere ab, was zu einem Übergang in einen linearen Anstieg des Brechungsindex führt. Insgesamt resultiert dies in einem S-förmigen Verlauf für den Brechungsindex.

Ein kritischer Vergleich zwischen den empirischen und simulierten Beugungsintensitäten offenbart ein unterschiedliches kinetisches Verhalten in den unbelichteten Bereichen. Das Fourier-Modell, das eine streng lokalisierte Polymerisation annimmt, spiegelt die Intensitäten der 0. und 1. Ordnung des Experiments für die erste Hälfte

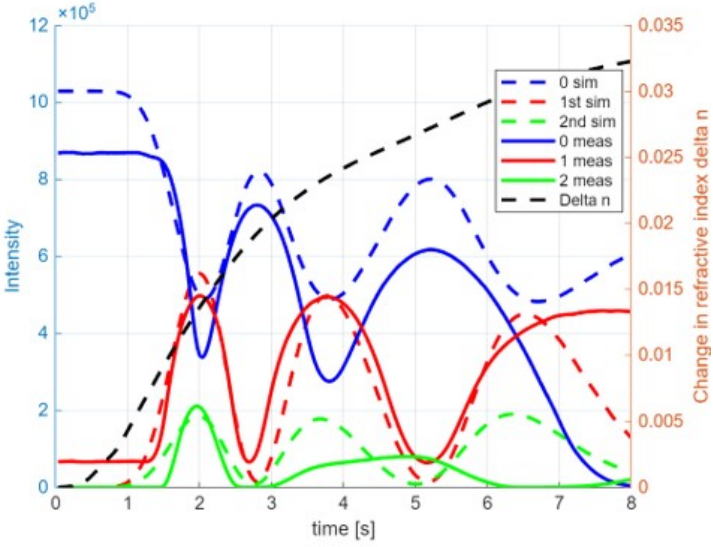


Abbildung 5: Vergleich der gemessenen Intensitäten mit den simulierten Intensitäten für die 0., 1. und 2. Beugungsordnung.

der Aushärtungsdauer ($t \leq 6$ s) perfekt wider und berücksichtigt die 3-Peak-Substruktur präzise als Artefakt der zentralen und peripheren Totzonen der Pixel [4].

Bei $t > 6$ s klingen die empirischen Beugungsordnungen jedoch deutlich langsamer ab, als es die idealisierte Simulation vorhersagt, während die 2. Ordnung bereits bei $t = 3$ s einen asymmetrischen Verlauf zeigt. Diese Abweichungen untermauern die Hypothese einer parasitären Aushärtung in den Dunkelzonen. Lichtstreuung über die Schichttiefe und die laterale Diffusion langlebiger, radikalisierter Oligomere initiieren eine verzögerte Polymerisation in den theoretisch unbelichteten Volumina.

Durch die numerische Einführung eines verzögerten Indexanstiegs für die Dunkelzonen (gestrichelte schwarze Linie in Abbildung 5) wurde ein erster Ansatz für ein angepasstes Modell geschaffen. Dadurch kann eine leichte Asymmetrie der simulierten Peaks (siehe 2. Ordnung) realisiert werden, was jedoch einen nicht realen Peak bei 6 s erzeugt.

5 Zusammenfassung

Diese Studie liefert umfassende empirische und simulative Beweise, die die Kinetik der Photopolymerisation direkt mit der Entwicklung der optischen Eigenschaften verknüpfen. Während die S-förmige zeitliche Entwicklung des Brechungsindex kontrastreiche Mikrooptiken ermöglicht, schränken parasitäre Phänomene wie Radikaldiffusion und Streuung die räumliche Auflösung stark ein. Um hochpräzise GRIN-Elemente zuverlässig zu drucken, müssen zukünftige Belichtungsstrategien eine adaptive Grauwertmodulation integrieren, um den verzögerten Indexanstieg in unbelichteten Bereichen aktiv zu kompensieren.

Danksagung

Die Autor möchte besonders Manuel Rank und Anne Harth für ihre wertvollen Beiträge danken, sowie der Deutschen Forschungsgemeinschaft für die Finanzierung (Fördernummer HE 3533/13-1).

Literatur

1. P. F. Jacobs, “Fundamentals of Stereolithography,” *1992 International Solid Freeform Fabrication Symposium*, 1992.
2. A. Heinrich, Ed., *3D Printing of Optical Components*, ser. Springer Series in Optical Sciences. Cham, Switzerland: Springer, 2021, no. volume 233.
3. S. Haegele, M. Rank, R. A. Terborg, A. Heinrich, and V. Pruneri, “Comprehensive optical monitoring of photopolymer curing for additive manufacturing of diffractive elements,” *Optics Express*, vol. 32, no. 20, p. 35212, Sep. 2024.
4. A. Heinrich and M. Rank, “Dynamic analysis of refractive index evolution and diffraction properties during single-photon polymerization of photopolymers for micro-optical applications,” *arXiv*, 2026.

Bildverarbeitung spielt in vielen Bereichen der Technik zur schnellen und berührungslosen Datenerfassung eine Schlüsselrolle. Beispielsweise in der Qualitätssicherung industrieller Produktionsprozesse, in der Robotik und zur Fahrerassistenz haben sich Bildverarbeitungssysteme einen unverzichtbaren Platz erobert. Diese Entwicklung wird unterstützt durch die Verfügbarkeit qualitativ hochwertiger und günstiger Sensorsysteme sowie durch die Zunahme der Leistungsfähigkeit von Rechnersystemen.

Der vorliegende Tagungsband des »Forums Bildverarbeitung / Image Processing Forum«, das am 5. und 6. November 2026 in Karlsruhe als gemeinsame Veranstaltung des Karlsruher Instituts für Technologie und des Fraunhofer-Instituts für Optronik, Systemtechnik und Bildauswertung stattfand, enthält die schriftlichen Aufsätze der eingegangenen Beiträge. Darin wird über aktuelle Trends und Lösungen der Bildverarbeitung in den methodischen Schwerpunkten **Bildgewinnung, Bildverarbeitung und Messtechnik, Robotik und Fahrzeuge, Maschinelles Lernen und künstliche Intelligenz sowie Qualitätssicherung** berichtet.

ISSN 2510-7224
ISBN 978-3-7315-1502-9

Gedruckt auf FSC-zertifiziertem Papier