Automatisierte Workflowanalyse im OP für die chirurgische Logistik

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für
Elektrotechnik und Informationstechnik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Lukas Kohout
geb. in Karlsruhe

Tag der mündlichen Prüfung:	10.04.2024
Hauptreferent:	Prof. Dr. rer. nat Wilhelm Stork
Korreferent:	Prof. Dr. med. Stephan Kruck

Zusammenfassung

Zur Optimierung der Logistik rund um den Operationsbereich in Krankenhäusern wird in dieser Dissertation untersucht, wie eine automatisierte Workflowanalyse von Operationen innerhalb eines intelligenten Operationssaals umgesetzt werden kann. Die Argumentation erfolgt dabei anhand der laparoskopischen Cholezystektomie als exemplarischer OP-Typ. Als Grundlage des entwickelten Konzepts dient eine detaillierte Analyse wiederkehrender Abläufe im OP-Umfeld im Allgemeinen sowie dem Verlauf und der eingesetzten Ressourcen bei der Cholezystektomie im Speziellen. Das erarbeitete Gesamtkonzept dieser Arbeit sieht ein modulares, multimodales Multitask-Netzwerk bestehend aus zunächst drei Teilsystemen für unterschiedliche Aspekte der Workflowanalyse vor. Die Teilsysteme analysieren dabei folgende Aspekte: die Aktivität im Operationssaal, die Nutzung der Materialien am Instrumententisch sowie die Erkennung einzelner OP-Phasen anhand des endoskopischen Kamerabildes. Die Umsetzung dieser Komponenten erfolgt aufgrund mangelnder zusammenhängender Trainingsdaten dabei zunächst unabhängig voneinander als eigenständige Einheiten.

Die Aktivitätsanalyse basiert auf der Auswertung von Gelenkpunkten, welche mittels Methoden des Posentrackings extrahiert werden. Durch Aggregation von Bewegungsvektoren zwischen einzelnen Gelenkpunkten über die Zeit können Aussagen über das Aktivitätslevel getroffen werden. Dabei ermöglichen unterschiedliche Filtermethoden große und kleine Bewegungen zu differenzieren und somit verschiedene Arten der Analyse. Im Rahmen dieses Experiments wird zusätzlich anhand eigens aufgezeichneter Videodaten die Re-Identifizierbarkeit von Personen in medizinischer Kleidung untersucht und Mitigationsstrategien zum Umgang mit der Verdeckung von diskriminierenden Merkmalen aufgezeigt.

Für die Detektion der zeitlichen Abfolge der genutzten OP-Instrumente werden insgesamt 26 verschiedene Modelle aus dem Stand der Technik mit eigenen Daten von realen Instrumenten mittels Transfer-Learning neu trainiert und bzgl. Nutzbarkeit im vorliegenden Kontext verglichen. Dabei wird deutlich, dass nur kleine Unterschiede in der Erkennungsqualität vorherrschen. Allerdings sind in Bezug auf den Ressourcenbedarf sowohl an die Hardware als auch an die Datenquantität signifikantere Unterschiede erkennbar. Letztendlich zeigen insbesondere verschiedene Varianten der YOLO-Familie, allen voran YOLOv5 und YOLOv8, Ergebnisse, die einen Einsatz im realen Umfeld ermöglichen. YOLOv8 in der m-Variante stellt dabei mit einer mAP und einem F1-Score von jeweils 99,5 % sowie einer Inferenzzeit von durchschnittlich weniger als 1,8 ms den besten Kompromiss zwischen Erkennungsqualität und Ressourcenbedarf dar.

Das Teilsystem zur OP-Phasenerkennung in Endoskopvideos nutzt ein Transformer-Modell, welches den zeitlichen Verlauf der Instrumentennutzung sowie ein Histogramm aller zuvor erkannten Phasen der laufenden Operation als Eingabedaten verarbeitet. Zur Umsetzung dieses Konzepts werden die öffentlichen Datensets Cholec80 und HeiChole genutzt. Auf den Cholec80-Daten kann eine Genauigkeit von 90 % und ein F1-Score von 75 % erreicht werden. Auf den komplexeren HeiChole-Daten ergibt sich ein durchschnittlicher F1-Score von ca. 53 %. Beides ist vergleichbar mit dem Stand der Technik, wobei die Modellkomplexität in der vorliegenden Arbeit geringer ist als bei den betrachteten Arbeiten anderer Forschungsgruppen. Im weiteren Verlauf kann das Modell bzgl. der HeiChole-Daten durch Kombination der beiden Datensets noch verbessert werden. Die Datenqualität und -quantität der verfügbaren Datensets ist für ein solches Deep-Learning-Modell allerdings nicht zufriedenstellend, so dass hier durch Optimierung der Trainingsdaten generell bessere Ergebnisse erwartet werden.

Die vorliegende Arbeit liefert somit einen Beitrag zu verschiedenen Aspekten der Workflowanalyse im OP-Umfeld und bildet die Grundlage für weitere Anwendungsmöglichkeiten von KI-Methoden im chirurgischen Kontext.

Abstract

In order to optimise the logistics around the operating theatre in hospitals, this dissertation examines how an automated workflow analysis of operations can be implemented within an intelligent operating room. The argumentation is based on laparoscopic cholecystectomy as an exemplary type of operation. The developed concept is based on a detailed analysis of recurring processes in the operating theatre environment in general and the course and resources used in cholecystectomy in particular. The overall developed concept for this work provides for a modular, multimodal multitask network consisting initially of three subsystems for different aspects of the workflow analysis. The subsystems analyse the following aspects: the activity inside the operating room, the use of materials on the instrument table and the recognition of individual surgical phases based on the endoscopic camera image. Due to a lack of coherent training data, these components are initially implemented independently of each other as separate units.

The activity analysis is based on the evaluation of keypoints representing joints, which are extracted using pose tracking methods. By aggregating movement vectors between individual keypoints over time, statements can be made about the activity level. Different filtering methods make it possible to differentiate between large and small movements and thus different types of analysis. As part of this experiment, the re-identifiability of persons in medical clothing is also investigated using specially recorded video data and mitigation strategies for dealing with the masking of discriminatory features are demonstrated.

For the detection of the temporal sequence of the surgical instruments used, a total of 26 different models from the state of the art are retrained with own

data from real instruments using transfer learning and compared with regard to usability in the present context. It becomes clear that there are only minor differences in the quality of recognition. However, more significant differences are recognisable in terms of the resources required for both the hardware and the quantity of data. Ultimately, different variants of the YOLO family in particular, especially YOLOv5 and YOLOv8, show results that allow them to be used in a real environment. With a mAP of 99.5 percent, a F1 score of 99.5 percent and an inference time of less than 1.8 ms on average, YOLOv8 in the m variant represents the best compromise between recognition quality and resource requirements.

The subsystem for OR phase recognition in endoscopic videos uses a transformer model that processes the temporal progression of instrument use and a histogram of all previously recognised phases of the current operation as input data. The public data sets Cholec80 and HeiChole are used to implement this concept. An accuracy of 90 percent and an F1 score of 75 percent can be achieved on the Cholec80 data. On the more complex HeiChole data, an average F1 score of approx. 53 percent is achieved. Both are comparable to the state of the art, although the model complexity in the present work is lower than in the considered work of other research groups. The model can be further improved with regard to the HeiChole data by combining the two data sets. However, the data quality and quantity of the available data sets is not satisfactory for such a deep learning model, so that better results are generally expected by optimising the training data.

This, this work makes a contribution to various aspects of workflow analysis in the operating theatre environment and forms the basis for further possible applications of AI methods in the surgical context.

Präambel

Nachfolgend sind weibliche, diverse und männliche Geschlechteridentitäten an jeder Stelle gleichermaßen angesprochen. Des Weiteren wird aufgrund von integrierten verlinkten Referenzen und zur besseren Lesbarkeit empfohlen, das Dokument in der farbigen Digitalversion zu lesen.

Danksagung

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am FZI Forschungszentrum Informatik. Mein besonderer Dank gilt Prof. Dr. rer. nat Wilhelm Stork für die Promotionsmöglichkeit und die Betreuung dieser Arbeit. Sein Blick auf Interdisziplinarität und praktische Anwendbarkeit neuester wissenschaftlicher Erkenntnisse waren eine große Inspiration für meine wissenschaftliche Arbeit sowie meine persönliche Entwicklung. Prof. Dr. med. Stephan Kruck danke ich für die Übernahme des Korreferats und das damit verbundene Engagement.

Für die umfangreichen fachlichen Diskussionen danke ich insbesondere Christoph Zimmermann, Marc Schroth und Matthias Diehl. Des Weiteren möchte ich bei meinen Kollegen und Freunden Dr.-Ing. Markus Schinle, Kai Zhou, Dr.-Ing. Jennifer Zeilfelder, Jacob Langner, Lennart Ries, Dr.-Ing. Timon Blöcher und Christina Erler für das kollegiale und stets inspirierende Arbeitsumfeld bedanken. Ich danke außerdem allen von mir betreuten Studierenden sowie Projektpartnern, insbesondere Matthias Lambertz und Jens Rennert, die durch ihre Arbeit wichtige Beiträge zu den vorliegenden Entwicklungen geleistet haben.

Ganz besonderer Dank gilt meiner Familie und Freunden, die für regelmäßigen Ausgleich zum beruflichen Alltag sorgen. Von ganzem Herzen danke ich meiner Frau Lisa für ihre Geduld und Unterstützung während ich in die Ausgestaltung und Fertigstellung meines Promotionsvorhabens vertieft war. Zuletzt danke ich Lucy und Wilma, die für die nötigen Bewegungspausen und Frischluftzufuhr gesorgt haben.

Karlsruhe, im Februar 2024 *Lukas Kohout*

Inhaltsverzeichnis

Abkürzungen und Symbole

AUC Area under the Curve

AP Average Precision

BMBF Bundesministerium für Bildung und Forschung

CEC Constant Error Carrousel

CNN Convolutional Neural Network

CT Computertomographie

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DRG diagnosebezogene Fallgruppe (engl. Diagnosis Related Group)

ERCP endoskopische retrograde Cholangiopankreatikographie

FN False Negative

FP False Positive

FPS Frames pro Sekunde

GRU Gated Reccurent Unit

GFLOPs Giga Floating Point Operations

GMACs Giga Multiply–Accumulate Operations

HAR Human Action Recognition

HF Hochfrequenz

HMM Hidden Markov Model

HOG Histogram of Oriented Gradients

HL7 Health Level 7

ID Identifikationsnummer

IoU Intersection over Union

IP Internet Protocol

JSON JavaScript Object Notation

KAS Klinisches Arbeitsplatzsystem

KI Künstliche Intelligenz

KIS Krankenhausinformationssystem

KRINKO Kommission für Krankenhaushygiene und Infektionsprävention

LSTM Long-Short-Term-Memory Netz

mAP Mean Average Precision

MDR Medical Device Regulation

MICCAI Medical Image Computing and Computer Assisted Intervention Society

WMBW Ministerium für Wirtschaft Arbeit und Wohnungsbau Baden-Württemberg

MPBetreibV Medizinproduktebetreiber-Verordnung

MRCP Magnetresonanzcholangiopankreatikographie

NAKI Nationaler Arbeitskreis zur Implementierung der EU-Verordnungen über Medizinprodukte und In-vitro-Diagnostika

NLP Natural Language Processing

NMS Non-Maximum Suppression

NOTES Natural Orifices Translumінal Endoscopic Surgery

OP Operation

OP-Saal Operationssaal

PACS Picture Archiving and Communication System

PDMS Patientendatenmanagementsystem

PKI Pre-Knowledge Inference

P-Kurve Precision-Kurve

PoE Power over Ethernet

PR-Kurve Precision-Recall-Kurve

R-CNN Region-based Convolutional Neural Network

RPN Region Proposal Network

Re-ID Re-Identifikation

RKI Robert Koch Institut

R-Kurve Recall-Kurve

RNN Rekurrentes Neuronales Netzwerk

ROI Region of Interest

RT-DETR Real-Time Detection Transformer

SDS Surgical Data Science

SIFT Scale-Invariant Feature Transform

SILC Single-Incision-Laparoscopic-Surgery

SPP Spatial Pyramid Pooling

SSD Single-Shot-Multibox-Detektor

SVM Support Vector Machine

TCN Temporal Convolutional Network

TN True Negative

TP True Positive

UML Unified Modeling Language

vdek Verband der Ersatzkassen e. V.

ViT Vision Transformer

YOLO You Only Look Once

1 Einleitung und Motivation

Die in Deutschland vorherrschende duale Krankenhausfinanzierung setzt sich aus Investitionen aus dem Budget der Bundesländer sowie Einnahmen über die Krankenkassen zusammen. Dabei soll der Anteil der Bundesländer die Kosten für die Aufrechterhaltung der Versorgungsstrukturen (z. B. Gebäude und Infrastruktur) decken, wohingegen Personal- und Betriebskosten durch den Krankenkassenanteil finanziert werden müssen [17]. Die Aufteilung verschiebt sich jedoch immer weiter in Richtung der beitragsfinanzierten Kostenträger, so dass inzwischen weniger als zehn Prozent des Gesamtbudgets eines Krankenhauses aus der Länderfinanzierung stammt. Der Verband der Ersatzkassen e. V. (vdek) zeigt in [156] sogar auf, dass die Investitionsmittel der Länder zwischen 1991 und 2017 um 18 Prozent abgebaut wurden, wohingegen im gleichen Zeitraum die Krankenhausausgaben von 29 Mrd. Euro auf insgesamt 75 Mrd. Euro angestiegen sind. Damit verringerte sich der Anteil der Krankenhausfinanzierung der Länder seit 1991 effektiv von über zehn Prozent auf weniger als vier Prozent. Die Finanzierung über den Krankenhausbetrieb wiederum ist maßgeblich geprägt durch *diagnosebezogene Fallgruppen (engl. Diagnosis Related Groups) (DRGs)*, ein System, das letztendlich das Leistungsvolumen, also die Quantität von Eingriffen, innerhalb eines Krankenhauses belohnt. Laut [119] basiert über 90 Prozent des Gesamtbudgets deutscher Krankenhäuser auf dieser Form der Vergütung. Dies zeigt, dass der OP-Bereich zu den Haupterlösquellen eines Krankenhauses gehört (vgl. [17], [90], [150], [158]), weshalb er auch als dessen „Motor“ [21], [90] oder „Herzstück“ [47] bezeichnet wird. Entsprechend haben Einschränkungen im OP-Betrieb direkten Einfluss auf die wirtschaftliche Situation einer Einrichtung und „vor dem Hintergrund der Unterfinanzierung

steigt bei den Krankenhäusern der Anreiz zur Leistungsausweitung, um die Kosten decken zu können“ [156].

Aus den genannten Gründen ist ein immer größer werdendes Ziel der Krankenhausbetreiber die Maximierung der Auslastung des OP-Bereichs und der damit verbundenen Ressourcen. Waeschle et al. resümieren in [158], dass die vorherrschenden Bedingungen „nur durch Erlössteigerung, rationalisierende Prozessoptimierung und Kostenminimierung kompensiert werden“ können. Auch Knauthe zeigt in [90] diverse Optimierungspotenziale durch bessere OP-Planung. So konnte er in seinen Analysen zu einem Beispielkrankenhaus im Jahr 2001 insgesamt Verschiebungen von Operationen (OPs) bei 383 Patientinnen und Patienten ausfindig machen, was in einer mittleren Wartedauer von bis zu 16,4 Tagen resultierte. Weiterhin wurden 12,1 Prozent der geplanten Leistungen in einen anderen Saal verschoben. Beide Werte weisen auf unzureichende Planung der Eingriffe hin. Die Leerlaufzeiten, also nichtproduktive Anteile der Bindungszeit im Operationssaal (OP-Saal), summiert sich laut dieser Studie auf fast vier Prozent, was ca. 78 Minuten pro OP-Saal pro Tag entspricht. Gleichzeitig war im Rahmen der Tagesplanung aufgefallen, „dass [. . .] nur 5,58 Prozent [der Pläne] mit einer optimalen OP-Auslastungszeit [. . .] zu verzeichnen waren“ [90]. Die Auswertungen ergaben weiterhin, dass durch organisatorische Optimierung der OP-Planung Überstunden massiv abgebaut und gleichzeitig fast sechs OPs mehr pro Tag durchgeführt werden können.

Abgesehen von diesen nicht ausgeschöpften Erlösmöglichkeiten, führt suboptimale Kapazitätsplanung im OP auch zu erhöhter Arbeitslast und Überstunden. Dies wiederum bedingt einerseits Spannungen im Team sowie reduzierte Motivation bei den Mitarbeitenden [150] und führt andererseits aber auch zu vermeidbaren Mehrkosten, da schätzungsweise 43 Prozent der direkt anfallenden Aufwendungen einer OP den Personalkosten zuzuordnen sind [90]. Eine Auswertung der OP-Barometer-Befragungen über zehn Jahre hinweg zeigen eine sinkende Tendenz in der Mitarbeitendenzufriedenheit, was unter anderem auch im schlechten Organisationsgrad der jeweiligen OP-Bereiche begründet liegt [21].

Zusammengefasst führt schlechte OP-Planung laut Tschudi et al. zu folgenden Nachteilen [150]:

- Es kommt zu vermeidbaren Über- und Unterauslastungen der OP-Säle mit Fallabsagen bzw. Erlösausfällen
- Zur Kompensation der fehlerhaften Kapazitätsallokation ist ein relevanter Koordinationsaufwand notwendig
- Das Risiko einer reduzierten Prozessqualität steigt
- Abgesagte Fälle und zusätzlich anfallende Überstunden führen zu Spannungen im OP-Team und reduzieren die Motivation der Mitarbeitenden.

Im Umkehrschluss bezeichnen die gleichen Wissenschaftler das OP-Kapazitätsmanagement als „eines der wichtigsten Instrumentarien zur Beeinflussung der Gesamtkosten der chirurgischen Leistungserstellung“ und führen folgende Ziele für dessen erfolgreiche Umsetzung auf:

- Sicherstellung der bedarfsgerechten Verfügbarkeit von Saalkapazitäten (auch für die Notfallversorgung)
- Maximierung der Saalauslastung
- Effizienter Ressourceneinsatz
- Minimierung anfallender Überstunden
- Minimierung der Ausfallquote elektiver Operationen und damit verbundene positive Effekte auf die Servicequalität.

Diese Aussage sowie die genannten Ziele werden u. A. in [10], [17], [47], [53], [80], [90], [150] und [158] bestätigt.

Eine große Herausforderung beim OP-Management ist die schwierige Planbarkeit von Eingriffen, bspw. aufgrund unvorhersehbarer Komplikationen und Verzögerungen oder zusätzlich eintreffender Notfallpatienten. Dies erschwert

eine adäquate Zeitabschätzung von OPs sowie die damit verbundene Personal- und Saalplanung. Mit steigender Komplexität und Dauer des Eingriffs wird zunehmend auch die Abschätzbarkeit zum Problem. Darüber hinaus erfolgen aktuelle Abwägungen üblicherweise nach subjektiven Kriterien und nicht individualisiert, bspw. bezugnehmend auf das OP-Team. Aus diesen Gründen wird üblicherweise zwischen geplanten Eingriffen ein Puffer eingeplant, der durch optimierte Schätzmethoden deutlich reduziert werden könnte. Insgesamt birgt die Thematik der Planbarkeit noch erhebliche Optimierungspotenziale.

Neben dem Bedarf das OP-Management zu optimieren, steigt gleichzeitig die Komplexität der eingesetzten Technologien innerhalb des OP-Saals. Durch die wachsende Anzahl an medizintechnischen Geräten, computer- und roboterbasierter Assistenzsystemen [89] und der damit einhergehenden uneinheitlichen und nicht intuitiven Bedienbarkeit, bedingt durch die Vielzahl einzelner Insellösungen, entstand in den 1990er Jahren der Bedarf einer Vernetzung und damit Vereinheitlichung der genutzten Geräte zum sog. *integrierten Operationssaal*. Dieser ermöglicht eine effiziente Bedienung, eine schnelle und einfache Datenerfassung, unkomplizierte Dokumentation sowie einen unbeschränkten Datenaustausch. Sowohl medizinische Geräte, Video- und Datenquellen als auch Peripheriegeräte sind auf den Bediener angepasst und aus dem Sterilbereich des OP-Saals steuerbar. Durch die beschriebenen Vereinheitlichungen auf verschiedenen Ebenen kann zusätzlich die Vernetzung und Einbindung von Daten in das Krankenhausinformationssystem (KIS) erzielt werden. Insgesamt sollen mittels des integrierten Operationssaals Medizingeräte vernetzt und die Effizienz eines Krankenhauses erheblich erhöht werden [134], [87], [132]. Der hohe Vernetzungsgrad eines vollständig integrierten OP-Saals und der gesammelten Informationen ermöglicht mit Hilfe moderner informationstechnischer Methoden, insb. aus dem Bereich der Datenanalyse, die Weiterentwicklung zum *intelligenten Operationssaal*, der eine systematische Auswertung aller verfügbaren Daten ermöglicht. Dadurch können bisher nicht erkennbare Zusammenhänge erfasst und neuartige Unterstützungsmöglichkeiten für die Patienten und das OP-Team umgesetzt werden.

Im Operationssaal der Zukunft entstehen somit neuartige, bisher ungenutzte Potenziale, u. A. bei der Planungsoptimierung, der automatisierten Dokumentation oder der Qualitätskontrolle während laufender OPs.

1.1 Motivation

Die Erläuterungen im vorherigen Abschnitt prognostizieren, dass durch Optimierungen im OP-Management Verbesserungspotenziale in der gesamten OP-Logistik, u. A. bestehend aus Patiententransport und -vorbereitung, Personalplanung, Materialverwaltung oder Terminierung der zur eigentlichen OP zugehörigen vor- und nachgelagerten Prozesse, ausgeschöpft werden können. Dies wiederum führt zu gesteigertem Patientenwohl, Entlastung des OP-Teams sowie Verbesserung der wirtschaftlichen Situation des Krankenhauses.

Folgende Hochrechnung verdeutlicht letzteren Punkt: Unter der Annahme eines mittleren Erlöses pro OP-Minute über alle Kostenarten und DRGs von ca. 17 Euro [158] besteht ein Steigerungspotenzial von mehr als 340.000 Euro pro Jahr pro Operationssaal. Bei 1.887 deutschen Krankenhäusern im Jahr 2021 [145], von denen, laut OP-Barometer der letzten Jahre, die Mehrzahl mindestens 4 OP-Säle besitzt [18], [19], [20], ergibt sich hierbei eine potenzielle Gesamtsumme an zusätzlichen Erlösen von mehr als 2,6 Mrd. Euro pro Jahr. Dabei sind die kumulativen Vorhaltekosten eines OP-Saals von ca. 800 Euro pro Stunde [80] noch nicht berücksichtigt. Darüber hinaus können noch zusätzliche Personalkosten aufgrund der vermeidbaren Mehrarbeit mit einberechnet werden. Laut Knauthe beziffern diese sich pro Person im Mittel auf ca. 0,60 Euro pro OP-Minute für den ärztlichen Dienst und ca. 0,40 Euro pro OP-Minute für den Funktionsdienst [90]. Dadurch entstehen, beim typischen Einsatz von drei Mitarbeitenden des ärztlichen Dienstes und vier Mitarbeitenden des Funktionsdienstes pro OP, Mehrkosten von mehr als 200 Euro pro Überstunde, was sich für das Krankenhaus im Beispiel aus [90] auf fast 78.000 Euro im Jahr summierte. Hochgerechnet auf alle 1.887 deutschen Krankenhäuser ergeben sich dadurch Gesamtkosten von mehr als 147 Mio. Euro pro Jahr. Da die hier

angesetzten Werte bereits aus dem Jahr 2001 stammen, sind aufgrund der Lohnentwicklungen der letzten 20 Jahre inzwischen sogar noch wesentlich höhere Beträge zu erwarten.

Für die Ausschöpfung dieser Potenziale ist einer der wichtigsten Bausteine im OP-Management die möglichst effiziente Auslastung aller vorhandenen Ressourcen. Dies beinhaltet auch eine zunehmend dynamische OP-Planung, basierend auf dem aktuellen Verlauf in allen OP-Sälen einer Einrichtung. Dazu erklären Tschudi et al. [150]: „Für eine effektive Kapazitätssteuerung im OP ist es wichtig, Informationen über [. . .] freiwerdende Kapazitäten so früh wie möglich zu erhalten“. Entsprechend ist auch eine adäquate, kontinuierliche Abschätzung der Restlaufzeit aller OPs erforderlich. Bisher erfolgt die Zeitabschätzung eigenverantwortlich und parallel zur laufenden Operation durch das Personal vor Ort. Dies ist in hohem Maße erfahrungsabhängig und folglich ungenau, insb. bei wenig eingespielten Teams. Außerdem führt die Abfrage über den aktuellen Status vom Planungsteam außerhalb des OP-Saals zu zusätzlichem kognitivem Stress, was sich vor allem in herausfordernden OP-Situationen negativ auf den Operationsverlauf auswirken kann. Eigene Beobachtungen und persönliche Gespräche mit Ärzten, Pflegepersonal und OP-Management haben bestätigt, dass Bedarf nach einer automatisierten Zeitabschätzung, ohne die Notwendigkeit der Interaktion mit dem OP-Team, besteht.

Ein weiterer Punkt, der sowohl Kostentreiber als auch Belastungsfaktor für das Personal darstellt, ist die Dokumentation behandlungsrelevanter Daten. Ziel der Dokumentation ist zum einen die Kontrolle durch den medizinischen Dienst für Haftungsfälle und zum anderen die Verfügbarmachung von Patienteninformationen und Transparenz über den Ressourceneinsatz. Eine Studie der HIMSS EUROPE aus dem Jahr 2015 zeigt allerdings auf, dass im Pflegedienst im Schnitt ca. 36 Prozent der Arbeitszeit für Dokumentationszwecke aufgewendet wird. Im ärztlichen Dienst sind es sogar 44 Prozent, wobei der Trend steigend ist [68]. Hinzu kommt, dass die Dokumentation nach wie vor häufig unstrukturiert stattfindet, was weitere Analysen der Inhalte, bspw. bzgl. Krankheitsverläufen von Patientinnen und Patienten, erheblich erschwert.

Die zuvor beschriebenen Entwicklungen hin zum intelligenten Operationssaal in Kombination mit der immer größeren Verbreitung von Methoden der künstlichen Intelligenz (KI) und des maschinellen Lernens, welche eine effiziente Auswertung großer Datenmengen erlauben, ermöglichen neuartige Konzepte, um diese Bedarfe zu adressieren. Bereits seit 2015 werden im Rahmen der *Endoscopic Vision Challenge* im 2-Jahres-Rhythmus verschiedene Fragestellungen im Bereich *Surgical Data Science (SDS)* bearbeitet [78]. Ein seit 2017 wiederkehrendes Thema dieser Challenges ist die chirurgische Workflowanalyse, bei der Handlungen und Prozessabläufe von chirurgischen Eingriffen automatisiert erfasst und für weitere Untersuchungen verarbeitet werden. Ein vielversprechender Ansatz dabei ist die Unterteilung in einzelne, spezifische OP-Phasen. Mithilfe von Vorwissen über den Prozess, wie bspw. typische Handlungssequenzen innerhalb der jeweiligen Abschnitte, lassen sich nach deren Erfassung Rückschlüsse über den Verlauf und somit auch auf die verbleibende Dauer ziehen. Zusätzlich können sämtliche Informationen, die erkannt und erfasst werden können, auch strukturiert und nachvollziehbar dokumentiert werden, was zu massiver Entlastung des medizinischen Personals führt.

Die bisher diskutierten Punkte tragen, neben den verbesserten wirtschaftlichen Aspekten, vor allem auch zur Verbesserung des Patientenwohls bei. Einerseits können die Abläufe der OP-Logistik optimiert werden, was dazu führt, dass diese für Patienten transparenter und Wartezeiten am Tag des Eingriffs minimiert werden. Durch Optimierung der Saalauslastungen und damit verbundener höherer OP-Dichte, werden aber auch Wartezeiten in der Anbahnung eines Eingriffs reduziert und somit schnellere Behandlungen ermöglicht. Dies wirkt sich wiederum positiv auf den Krankheitsverlauf der einzelnen Patienten aus. Gleichzeitig führt die Verkürzung der Wartelisten auch insgesamt zu Entlastungen im gesamten Gesundheitssystem, bspw. durch die Reduzierung der reinen Symptombehandlung aufgrund der Verzögerungen im Behandlungsverlauf.

In bisherigen Arbeiten lag der Fokus auf einzelnen Teilaspekten einer Operation, bspw. dem laparoskopischen Anteil, oder dem Einbezug einzelner Teilsysteme des integrierten OP-Saals, wie z. B. dem endoskopischen Kamerabild. Eine

multimodale Betrachtung des kompletten intraoperativen Prozesses unter Einbezug verschiedener Informationsquellen ist zum Zeitpunkt der Erstellung dieser Arbeit nicht bekannt. Aufgrund dieser motivierenden Faktoren sollen in der vorliegenden Arbeit darauf aufbauende Konzepte erarbeitet und bzgl. Umsetzbarkeit und Einsatzmöglichkeiten in der Praxis bewertet werden.

1.2 Ziel der Arbeit

Aus den in der Einleitung und Motivation genannten Argumenten ist das Ziel dieser Arbeit die Erforschung von Möglichkeiten für die automatisierte Workflowerkennung im OP zur Entlastung des Personals und Optimierung der zugehörigen logistischen Prozesse. Abgrenzend zu bestehenden Arbeiten auf diesem Gebiet soll der Fokus nicht auf einem speziellen Teilabschnitt einer OP oder einer spezifischen Informationsquelle liegen, sondern alle Abschnitte des perioperativen Prozesses innerhalb des OP-Saals mittels vernetzter, multimodaler Sensorik mit in die Betrachtung einbeziehen. Die folgende zentrale Forschungsfrage bildet dabei den Kern dieser Arbeit:

> *Wie können komplexe Handlungsabläufe und Prozesse zwischen mehreren Akteuren zuverlässig auf Basis multisensorieller Daten erfasst und analysiert werden?*

Basierend darauf ergeben sich, unter Beachtung der spezifischen Rahmenbedingungen des Anwendungsfalls, folgende unterstützende Teilfragen für die Beantwortung der Leitfrage:

1. Welche Anforderungen bestehen an ein Kontext-Erfassungssystem im regulierten OP-Umfeld?
2. Welche Sensorik ist für eine robuste Erfassung geeignet und wie kann diese zu einem Gesamtsystem integriert werden?
3. Welche Methoden und Konzepte sind notwendig, um eine Erfassung von OP-Phasen zu realisieren?

4. Wie können die daraus abgeleiteten Informationen zur Verbesserung der Arbeitsumgebung im OP sowie der Optimierung dessen Logistik genutzt werden?

1.3 Wissenschaftliches Umfeld

Wesentliche Teile dieser Arbeit entstanden im Rahmen des Forschungsprojekts „KIMONO - KI-getriebene Erkennung und Analyse von Handlungen und Prozessen auf Basis fusionierter Multisensordaten zur Optimierung von OP-Abläufen“, welches von Juni 2019 bis November 2023 von der Richard und Annemarie Wolf-Stiftung gefördert wurde. Ziel dieses Projekts ist die Erforschung geeigneter Sensorsettings und Methoden zur Erfassung von Prozessen im OP-Umfeld.

Weiterhin flossen Erkenntnisse aus den Projekten „HEIKE - IT-Assistenzsystem zur Verbesserung der Händedesinfektion in deutschen Krankenhäusern“, gefördert zwischen 06/2016 und 06/2019 vom Ministerium für Wirtschaft Arbeit und Wohnungsbau Baden-Württemberg (WMBW) (seit 2021 Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg) und „situCare - Situative Unterstützung und Krisenintervention in der Pflege“, gefördert zwischen 04/2016 und 06/2019 vom Bundesministerium für Bildung und Forschung (BMBF), mit in die Erarbeitung der vorliegenden Ausführungen ein.

1.4 Ablauf der Arbeit

Die Aktivitäten zur Anfertigung der vorliegenden Arbeit werden in insgesamt 8 Kapiteln dargestellt. Nach den Erläuterungen in Einleitung und Motivation werden in Kapitel 2 die Grundlagen zum Verständnis der medizinischen Aspekte sowie technische Hintergründe insb. bzgl. Kontexterfassung näher betrachtet. Darauf aufbauend wird in Kapitel 3 der Stand der Technik auf dem Gebiet der verschiedenen Methoden zur Erfassung von Kontexten, unterteilt nach den Anwendungsfällen der Erkennung von Objekten, Personen und Posen. Außerdem

wird im Unterkapitel 3.3 konkret der Stand der Technik zur Workflowanalyse im OP veranschaulicht, bevor in Unterkapitel 3.4 Datensätze für die datengetriebene Entwicklung von Algorithmen erörtert werden. Basierend auf diesen Erkenntnissen werden in Kapitel 4 zunächst Anforderungen an ein Kontexterkennungssystem im OP-Umfeld erarbeitet und die in den Grundlagen dargestellten medizinischen Abläufe technisch analysiert. Anschließend wird das Gesamtkonzept entwickelt und danach schließlich mögliche Risiken eines solchen Systems erarbeitet und diskutiert und das aus diesen Vorarbeiten abgeleitete Vorgehen zu deren Umsetzung dargestellt wird. Die Kapitel 5 bis 7 zeigen jeweils die Detailkonzeption, Umsetzung und Ergebnisdiskussion für die drei Teilsysteme zur Erfassung verschiedener Kontexte im OP. Abschließend fasst Kapitel 8 die Arbeit zusammen und bietet einen Ausblick auf mögliche Weiterentwicklungen und den praktischen Einsatz der vorgestellten Arbeiten.

2 Grundlagen

Dieses Kapitel erläutert die wesentlichen Grundlagen zum Verständnis der durchgeführten Arbeiten. Die Interdisziplinarität der zu bearbeitenden Fragestellung bedingt die Aufteilung in medizinische und informationstechnische Themen. Dafür werden zunächst notwendige Kenntnisse zum Krankenhausumfeld und der relevanten OP-Prozesse diskutiert. Der Schwerpunkt liegt dabei auf der Cholezystektomie, also der chirurgischen Entfernung der Gallenblase, als Beispieloperation. Anschließend werden die Grundlagen zur sensorischen Kontexterfassung dargestellt. Diese sind unterteilt in die geeignete Sensorik, Erkennungsmethoden und Möglichkeiten zur Evaluation der umgesetzten Systeme.

2.1 Abläufe und Prozesse im OP-Umfeld

Die Ziele der vorliegenden Arbeit erfordern ein umfangreiches Verständnis der Abläufe im Krankenhaus und speziell im OP-Saal. Dazu zählen sowohl die Prozesse und Abläufe während des chirurgischen Eingriffes im OP-Saal selbst als auch die Verfahrensweise vor und nach einer Operation. Im Folgenden werden die einzelnen Phasen sowie weitere relevante Inhalte im OP-Umfeld näher erläutert.

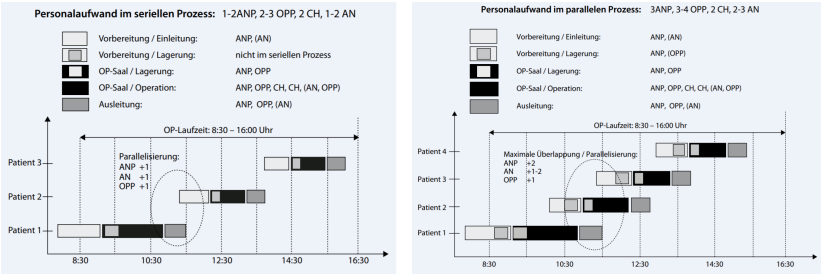
2.1.1 Der perioperative Prozess

Der gesamte Verlauf rund um einen operativen Eingriff wird als perioperativer Prozess bezeichnet. „Eingeschlossen ist [hierbei] die Zeit vor der Operation (präoperativ), die Zeit während der Operation (intraoperativ) und die Zeit nach dem Eingriff (postoperativ)“ [123]. Im Allgemeinen werden unter der präoperativen Phase alle Handlungen und Abläufe verstanden, die mit der Vorbereitung des Patienten außerhalb des OP-Bereiches zu tun haben. Dazu zählen unter anderem die Aufnahme des Patienten, das Patientengespräch mit dem Chirurgen und das Prämedikationsgespräch mit dem Anästhesisten, der Transport des Patienten sowie die Vorbereitung der Instrumente und des OP-Saals. Des Weiteren werden der präoperativen Phase die Betreuung des Patienten im Holding Room und die Vorbereitung der Anästhesie zugeordnet. Die intraoperative Phase schließt alle Handlungsabläufe vom Beginn der Einschleusung bis zum Ende des Ausschleusens des Patienten ein. Demnach sind die Prozesse der Übernahme des Patienten, die Umlagerung des Patienten auf den OP-Tisch, die Narkoseeinleitung und -ausleitung, die Überwachung der Narkose und alle OP-Schritte während der gesamten Schnitt-Naht-Zeit mit eingebunden. Die übrigen Behandlungsschritte im perioperativen Prozess sind der postoperativen Phase zuzuordnen. Dabei wird der Patient an das Personal im Aufwachraum übergeben, der OP-Saal wird gereinigt und zuletzt findet die Übergabe des Patienten an die nachgelagerte Station statt (vgl. [55]).

In allen drei Phasen arbeiten Vertreter unterschiedlicher Berufsgruppen sowohl seriell als auch parallel zusammen, wobei verschiedene Entscheidungskompetenzen vorliegen (vgl. [55]). Die einzelnen Berufsgruppen, speziell in Hinsicht auf die laparoskopische Cholezystektomie, werden in Abschnitt 2.1.5 näher aufgeführt. Im Folgenden wird zunächst die Differenzierung des parallelen und seriellen Operationsverlaufes näher betrachtet.

Eine erhebliche Herausforderung im perioperativen Operationsverlauf stellen Wartezeiten dar. Zur Vermeidung damit verbundener Probleme wird eine strukturierte Planung der Auslastung der OP-Säle und der benötigten Kapazitäten bzw. des Personalaufwandes für jede Operation angestrebt. Bei dieser Planung

wird zwischen seriellen und den parallelen Operationsverläufen unterschieden. Bei den seriellen Operationsverläufen wird nach einer OP der nachfolgende Patient erst dann vorbereitet und eingeleitet, wenn der vorhergehende Patient ausgeleitet wird. Somit läuft lediglich die Ausleitung des ersten Patienten und die Einleitung des zweiten Patienten parallel ab, alle weiteren Operationsabläufe sind seriell angeordnet. Dies hat zum Vorteil, dass weniger Personal benötigt wird, jedoch sind folglich auch weniger Operationen pro Tag möglich (vgl. [74]). Im Vergleich dazu liegen bei dem parallelen Operationsverlauf überlappende Anästhesieeinleitungen vor. Während der erste Patient also noch operiert wird, wird beim nachfolgenden Patienten bereits die Narkose eingeleitet und die Lagerung weitestgehend vorbereitet. Sobald der erste Patient den OP-Saal verlässt, kann der nächste Patient in den OP-Saal gebracht werden. Zwar ist für die parallele Ablaufplanung mehr Personal notwendig, allerdings sind mit dieser Organisation mehr Operationen pro Tag durchführbar, als bei dem seriellen Operationsverlauf und Wartezeiten können erheblich verkürzt werden (vgl. [74]). Abbildung 2.1 skizziert die unterschiedlichen Ablaufschemata.



(a) Seriell **(b)** Parallel

Abbildung 2.1: Serieller vs. paralleler Operationsverlauf nach [74]

Die Untersuchungen der vorliegenden Arbeit finden stets bezugnehmend auf die laparoskopische Cholezystektomie als Beispiel-OP statt. Dies ist zum einen darin begründet, dass die Cholezystektomie eine häufig durchgeführte OP ist und entsprechend als repräsentativ gelten kann (vgl. Abschnitt 2.1.2.) Zum

anderen ist sie in der Literatur bzgl. der Themen Workflowanalyse und OP-Phasenerkennung weit verbreitet. Dadurch ist eine Vergleichbarkeit mit anderen Forschungsgruppen möglich. Aus den genannten Gründen erfolgt in dieser Arbeit die Darstellung der grundlegenden Rahmenbedingungen und Abläufe im OP-Saal am Beispiel dieser spezifischen OP-Art. Bevor nachfolgend der perioperative Operationsverlauf der laparoskopischen Cholezystektomie näher erläutert wird, werden einige medizinische und statistische Grundlagen zur Cholezystektomie sowie die relevantesten Operationstechniken diskutiert.

2.1.2 Medizinische Grundlagen und Statistiken zur Cholezystektomie

Bei der Cholezystektomie handelt es sich um die chirurgische Entfernung der Gallenblase. Die Gallenblase ist ein Hohlorgan, welches an der Leberinnenfläche liegt (siehe Abbildung 2.2) und für die Speicherung der in der Leber produzierten Galle verantwortlich ist. Mit ihrem Fassungsvermögen von etwa 50 Millilitern bestehen weitere Funktionen der Gallenblase in der Konzentration der Galle, im Ausgleich von Druckunterschieden in den Gallengängen und in der Resorption von Fetten aus dem Dünndarm (vgl. [138]).

Für eine operative Gallenblasenentfernung können unterschiedliche Indikationen vorliegen. Dabei wird grundsätzlich zwischen relativen und absoluten Operationsindikationen differenziert. Gallenblasenpolypen, Gallenblasenkinesien, Gallenblasenpapillomatosen, sowie jede symptomatische Cholezystolithiasis (Gallensteinleiden) zählen zu den relativen Operationsindikationen. Absolute Operationsindikationen hingegen sind die akute Cholezystitis, die freie Gallenblasenperforation und das Gallenblasenempyem (vgl. [65]). Eine der häufigsten Indikationen ist dabei eine symptomatische Cholezystolithiasis (Gallensteine), die oftmals zu einer akuten Cholezystitis (Entzündung der Gallenblase) aufgrund eines temporären oder permanenten Verschlusses des Gallenblasenausführungsganges führt.

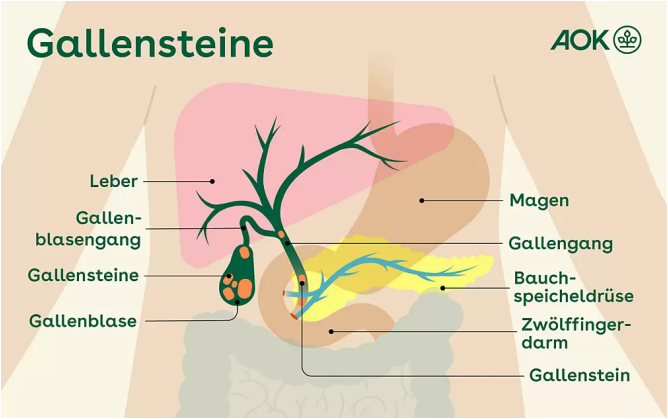


Abbildung 2.2: Darstellung der Lage der Gallenblase im Bauchraum (Quelle: [6])

Laut fallpauschalenbezogener Krankenhausstatistik gehört die Cholezystektomie mit jährlich ungefähr 200.000 Fällen zu den zehn am häufigsten durchgeführten OPs in Deutschland [144]. Dies spiegelt sich auch in der Prävalenz bei Gallensteinen von 15 - 20 Prozent innerhalb der deutschen Bevölkerung wieder [157]. Dabei ist die Mehrheit der Betroffenen weiblich (ca. 60 Prozent) und älter als 40 Jahre (ca. 83 Prozent) [77].

Die präoperative Diagnostik erfolgt klassisch mittels transabdomineller Sonographie [34]. Dies ist unter anderem in der hohen diagnostischen Genauigkeit begründet. Ahmed und Diggory zeigen in einer retrospektiven Analyse von 2100 laparoskopischen Cholezystektomien eine Sensitivität von 85 Prozent und eine Spezifität von 100 Prozent bei der sonographischen Identifikation von Gallensteinen [4]. Bei symptomatischen Patienten mit unauffälliger oder unklarer transabdomineller Sonographie kommt üblicherweise zur weiteren Abklärung entweder die endoskopische Sonographie oder die Magnetresonanzcholangiopankreatikographie (MRCP) zur Anwendung. In [56] werden beide Verfahren als gleichwertig bzgl. Sensitivität und Spezifität beurteilt. Insbesondere in der Komplikationsdiagnostik kann auch die Computertomographie (CT) eingesetzt werden [34]. Die *endoskopische retrograde Cholangiopankreatikographie (ERCP)* wird laut [16] aufgrund ihres Komplikationsprofils (invasives

Verfahren, Notwendigkeit von Sedierung und Kontrastmittel, Röntgenbelastung, Risiko einer Post-ERCP-Pankreatitis [41]) trotz ihrer hohen Genauigkeit bei der Steindetektion heutzutage nur noch selten zur Diagnostik auf diesem Gebiet eingesetzt. Vorteil dieser Methode ist jedoch, dass es sich hierbei nicht nur um ein diagnostisches Verfahren handelt, sondern bei der Untersuchung auch direkt therapeutische Maßnahmen getroffen werden können. Laborchemische Untersuchungen sind nur bei symptomatischer Cholezystolithiasis zielführend und entsprechend nur in Kombination mit den zuvor genannten Methoden zur Diagnostik geeignet.

Für die Entfernung der Gallenblase können verschiedene Operationsverfahren angewandt werden. Bei der konventionellen Cholezystektomie handelt es sich um die offene, chirurgische Entfernung der Gallenblase über einen Bauchschnitt, welcher üblicherweise entlang des rechten Rippenbogenrandes gesetzt wird. Da die konventionelle Cholezystektomie heute überwiegend durch das minimalinvasive laparoskopische Operationsverfahren abgelöst wurde, wird die offene Cholezystektomie lediglich bei operativen Komplikationen oder bei schwierigen pathoanatomischen Verhältnissen angewendet. Als heutiges Standardverfahren kommt in mehr als 90 Prozent der Fälle die laparoskopische Cholezystektomie zum Einsatz [77]. Hierbei handelt es sich weltweit um die am häufigsten durchgeführte viszeralchirurgische Operation (vgl. [37]). Die Prozedur erfolgt dabei „durch minimalinvasive Chirurgie in Form von wenigen Hautinzisionen über Trokare“ ([37]). In der Praxis sind unterschiedliche Formen der laparoskopischen Cholezystektomie zu finden. Die konventionelle laparoskopische Technik erfolgt über drei bis vier kleine Hautschnitte, durch die der Zugang zum Bauchinnenraum mittels der Trokare gelegt wird. Durch die stetige Weiterentwicklung neuer Technologien und Operationstechniken entstanden weitere Verfahren, wie die Natural Orifices Transluminal Endoscopic Surgery (NOTES) und die Single-Incision-Laparoscopic-Surgery (SILC). Bei der NOTES Technik wird der Operationszugang über natürliche Körperöffnungen, wie bspw. die Vagina, gelegt. Das spezielle Vorgehen der SILC Technik ermöglicht den chirurgischen Eingriff über einen zentralen Zugang, in den mehrere Instrumente eingeführt werden können (vgl. [36]).

Der Fokus dieser Arbeit liegt auf dem Standardverfahren der konventionellen laparoskopischen Cholezystektomie.

2.1.3 Operationsablauf der laparoskopischen Cholezystektomie

Als Laparoskopie wird die minimalinvasive Untersuchung der Bauchhöhle mittels Spezialendoskopen (sog. Laparoskope) bezeichnet. Dabei bleibt die Bauchdecke, abgesehen von kleinen Durchstichöffnungen, intakt [35], was zu weniger Komplikationen und besseren Heilungsaussichten führt. Die Zugänge in den Bauchinnenraum erfolgen über sog. Trokare und dienen der Einführung der benötigten Instrumente. Das laparoskopische Operationsverfahren wird einerseits für Operationen aus der Viszeralchirurgie, wie zum Beispiel der laparoskopischen Cholezystektomie, angewandt, andererseits aber auch für gynäkologische und urologische Operationen. Dabei bringt das laparoskopische Operationsverfahren im Vergleich zur offenen Chirurgie viele Vorteile mit sich. So werden zum einen die postoperativen Schmerzen reduziert, woraus sich zugleich eine frühere Mobilisierung und eine Verkürzung des Krankenhausaufenthaltes ergibt. Zum anderen liegt durch die laparoskopische Chirurgie ein verbessertes kosmetisches Ergebnis aufgrund kleinerer Narben vor. Dennoch gilt es zusätzlich die Nachteile der Laparoskopie zu beachten: So ist bspw. aufgrund der speziellen laparoskopischen Instrumente der technische Aufwand erheblich erhöht (siehe auch Abschnitt 2.1.6). Darüber hinaus ist der Umgang mit den Instrumenten und die Behebung von Komplikationen aufgrund der nur kleinen Zugänge in den Bauchinnenraum erschwert (vgl. [5]).

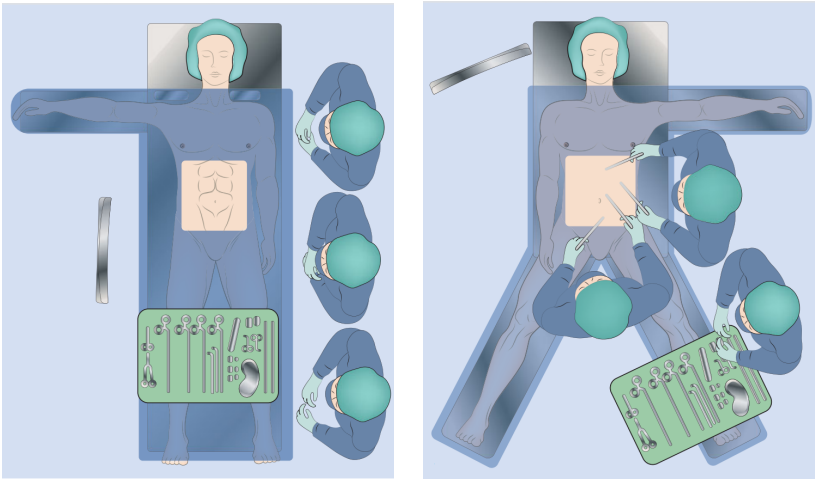
Da es sich bei der laparoskopischen Cholezystektomie um einen standardisierten chirurgischen Eingriff handelt, lässt sich der Operationsverlauf mit seinen einzelnen Prozessen in die wie bereits in Abschnitt 2.1 erläuterten präoperativen, intraoperativen und postoperativen Phasen einteilen. Im Folgenden wird, angelehnt an [8], der genaue Operationsverlauf beschrieben und dessen einzelne Prozesse anhand der drei Phasen differenziert.

2.1.3.1 Präoperativer Operationsverlauf

Zur Operationsvorbereitung gehört die Untersuchung des Bauchinnenraums und speziell der Oberbauchorgane. Das Standardverfahren dieser Voruntersuchung ist, wie bereits zuvor erläutert, die Sonographie (vgl. [25]). Zur weiteren Operationsvorbereitung gehört zudem die Aufklärung des Patienten. Der Arzt ist dazu verpflichtet, den Patienten über den Verlauf und die Risiken zu informieren. Insbesondere die Möglichkeit einer Konversion mit Laparotomie, also ein Wechsel zum offenen operativen Vorgehen im Falle von intraoperativen Komplikationen, ist Bestandteil der Aufklärung. Aufgrund der zuvor genannten Vorteile ist die laparoskopische Cholezystektomie sowohl bei Chirurgen als auch bei Patienten die präferierte OP-Technik. Darüber hinaus ist es bei elektiven Eingriffen erlaubt am Tag vor dem OP-Termin bis zum Abend normale Mahlzeiten zu sich zu nehmen und bis zu zwei Stunden vor dem chirurgischen Eingriff Wasser zu trinken (vgl. [25]).

Ist der OP-Saal frei, wird der Patient zum OP bestellt, von einer Stationspflegekraft vorbereitet und in seinem Stationsbett zum OP gebracht. Vor der Schleuse übergibt die Stationspflegekraft den Patienten an die OP-Pflegekraft, die sich um die Einschleusung des Patienten kümmert. In der Schleuse wird der Patient von seinem Stationsbett auf den OP-Tisch umgelagert. Anschließend wird der Patient auf dem OP-Tisch in den Einleitungsraum gebracht. Dort legt eine Anästhesiepflegekraft den venösen Zugang, kontrolliert die Vollständigkeit der Befunde und der Patient wird intubiert. Sofern alle Maßnahmen für die Einleitung des Patienten abgeschlossen sind, kann der Patient in den OP-Saal geschoben werden. Dort beginnen die letzten chirurgischen Vorbereitungen am Patienten: Die späteren Schnittstellen des Patienten werden desinfiziert und der Patient wird vom OP-Personal mit sterilen OP-Tüchern abgedeckt, wobei lediglich die Schnittstellen unverdeckt bleiben. Außerdem erfolgt die für den weiteren OP-Verlauf benötigte Lagerung des Patienten, welche für jeden chirurgischen Eingriff bestimmten Anforderungen folgt. Bei der laparoskopischen Cholezystektomie werden grundsätzlich zwei unterschiedliche Verfahren angewendet: die einfache Rückenlagerung oder die French Position. Bei der

einfachen Rückenlagerung wird der rechte Arm des Patienten im rechten Winkel ausgelagert. Außerdem befinden sich hierbei sowohl der Operateur, die Assistenzkraft, als auch die instrumentierende Pflegekraft auf der linken Körperseite des auf dem OP-Tisch liegenden Patienten (vgl. Abbildung 2.3a). Auch bei der French Position wird der Patient in eine Rückenlagerung versetzt. Allerdings werden seine Beine abgewinkelt und gespreizt positioniert, sodass der Operateur zwischen diesen steht. Die Assistenzkraft und die instrumentierende Pflegekraft befinden sich auf der linken Seite des Patienten, wobei die Assistenzkraft die Kamera hält und führt und die instrumentierende Pflegekraft dem Chirurgen die Instrumente reicht und abnimmt (vgl. Abbildung 2.3b). Bei dieser Lagerungsart kann entweder der rechte oder der linke Arm des Patienten ausgelagert werden. Für beide Lagerungsarten gilt, dass der Patient zusätzlich in die Fußtief- und Linkseitenlagerung, bzw. in die *Anti-Trendelenburg* und Linkseitenlagerung, gebracht wird. Dadurch werden naheliegende Organe aus dem Blickfeld im Arbeitsraum verschoben (vgl. [62]).



(a) Einfache Rückenlagerung **(b)** French Position

Abbildung 2.3: Lagerung des Patienten während der Cholezystektomie [62]

Zum präoperativen Prozess der laparoskopischen Cholezystektomie gehören neben den Vorbereitungen am Patienten zusätzlich das Richten der benötigten Instrumente und Materialien sowie des Laparoskopieturms. Details hierzu werden in Abschnitt 2.1.6 erläutert.

2.1.3.2 Intraoperativer Operationsverlauf

Nachdem der Patient und die Patientenumgebung entsprechend vorbereitet wurden, beginnt der intraoperative Operationsabschnitt. Bei der laparoskopischen Cholezystektomie wird der Zugang zum Bauchinnenraum in der Regel über vier Trokare gelegt. Dabei handelt es sich um drei Arbeitstrokare, durch die die laparoskopischen Instrumente eingeführt werden, sowie ein Optiktrokar, welcher der Einführung der Optik dient. Dadurch wird die Sicht in den Bauchinnenraum ermöglicht. Die Platzierung der Trokare ist in Abbildung 2.4 skizziert. Der Optiktrokar wird subumbilikal, unter dem Bauchnabel, platziert, während zwei Arbeitstrokare unterhalb des rechten Rippenbogens und ein Arbeitstrokar im linken Mittelbauch platziert werden. Die Arbeitstrokare, die am Rippenbogen platziert werden, verfügen üblicherweise über einen Durchmesser von fünf Millimetern. Der Trokar im linken Mittelbauch hingegen ist mit einem Durchmesser von zehn Millimetern für größere Instrumente geeignet (vgl. [25]). Alle Instrumente können auf diese Weise in einer entspannten Armhaltung von Chirurgen und Assistenten geführt werden und treffen im rechten Winkel auf die zu präparierenden Strukturen und Gewebe.

Vor dem Einführen des Optiktrokars erfolgt ein subumbilikaler Schnitt am Nabelrand. Danach wird zum Anlegen des Pneumoperitoneums Kohlenstoffdioxid mittels der Veressnadel und der Veresskanüle in den Bauchinnenraum des Patienten insuffliert. Dadurch hebt sich die Bauchdecke des Patienten und schafft Raum zum Operieren. Mit dem Einführen des Optiktrokars beginnt die diagnostische Laparoskopie. Hierfür wird für eine bessere Sicht auf den Monitoren der Raum stark abgedunkelt. Messungen während einer Hospitation einer realen Operation ergaben hier eine Beleuchtungsstärke im Raum von

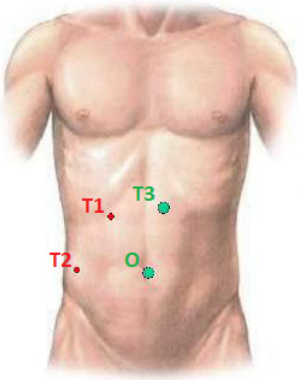


Abbildung 2.4: Platzierung der Trokare nach [25]. Die rot markierten Öffnungen für die Arbeitstrokare „T1“ und „T2“ haben meist einen Durchmesser von 5 mm, die grün markierten für Trokar „T3“ und den Optiktrokar „O“ üblicherweise einen Durchmesser von 10 mm.

lediglich 4 Lux. Nach der Inspektion der Bauchhöhle können die laparoskopischen Instrumente durch die Trokare in den Bauchinnenraum eingeführt werden. Dann wird zuerst die Gallenblase mit einer Fasszange über einen Arbeitstrokar, der am rechten Rippenbogen platziert ist, gefasst und über den Leberlappen geschoben. Sofern die Gallenblase mit fettreichen Adhäsionen angereichert ist, werden diese zunächst entfernt (vgl. [25]). Somit kann die Gallenblase mit der Fasszange weiter nach oben geschoben werden, mit dem Ziel, dass sich die Strukturen des Calot-Dreiecks anspannen. Der Ductus cysticus und die Arteria cystica werden dargestellt und freipräpariert. Der danach notwendige Verschluss der Arterien erfolgt über Clips, die mit einem Scharnier- und Schnappmechanismus versehen sind. Erst nach dem Anbringen der Clips wird sowohl der Ductus cysticus als auch die Arteria cystica mit einer Schere durchtrennt (vgl. [25]). Damit die Gallenblase aus dem Gallenblasenbett herausgelöst werden kann, erfolgt die Ausschälung mittels einer Schere und einem Ultraschallhaken. Dazu ist es von Bedeutung, dass die Gallenblase stets angespannt ist. Mögliche auftretende Blutungen aus dem Gallenblasenbett werden bipolar koaguliert. Nach der erfolgreichen Exzision der Gallenblase kann die Bergung aus dem

Bauchinnenraum stattfinden. Hierzu wird die Gallenblase mit der Fasszange am Infundibulum gefasst und über die Inzision mit zehn Millimetern Durchmesser im linken Mittelbauch aus der Bauchhöhle extrahiert (vgl. [25]). Die folgende Spülung des rechten Oberbauches und die Koagulation von restlichen Blutungen bilden den letzten chirurgischen Schritt innerhalb des Bauchraumes. Nach einer finalen Untersuchung und Kontrolle werden alle Instrumente und Trokare entfernt, sowie das Pneumoperitoneum abgelassen (vgl. [25]). Der schließlich letzte intraoperative Schritt ist die Hautnaht, bei der alle Inzisionen vom Operateur verschlossen werden.

2.1.3.3 Postoperativer Operationsverlauf

Sind alle operativen Maßnahmen am Patienten abgeschlossen, beginnt die postoperative Phase. Bereits kurz vor Ende der Hautnaht der Inzisionen wird die Narkose des Patienten ausgeleitet. Sobald die Operation vollständig abgeschlossen ist, wird der Patient extubiert sowie alle angeschlossenen Geräte entfernt. Zusätzlich wird der Patient auf dem OP-Tisch in die horizontale Ausgangslage gebracht, sodass er aus dem OP-Saal in den Ausleitungsraum geschoben werden kann. Beim Ausschleusen wird der Patient zurück auf das Stationsbett umgelagert. Danach wird er in seinem Stationsbett in den Aufwachraum gefahren, bevor er von einer Stationspflegekraft zurück auf die Station gebracht wird (vgl. [95]). Der gesamte Transportweg vom OP-Saal bis hin zum Aufwachraum wird vom Anästhesisten begleitet. Dieser informiert zusätzlich die Anästhesiepflegekraft im Aufwachraum über den Operationsverlauf, Besonderheiten und gegebenenfalls aufgetretene Komplikationen während der OP. Darüber hinaus gibt der Anästhesist der Anästhesiepflegekraft Anweisungen zur Medikation, bevor er die Einleitung des nächsten Patienten übernimmt (vgl. [95]).

Des Weiteren ist in die postoperative Phase zusätzlich die Deinstallation der Instrumente und der OP-Geräte mit eingebunden (vgl. [93]). Damit der OP-Saal schließlich für darauffolgende Operationen genutzt werden kann, erfolgt abschließend dessen Reinigung (vgl. [135]).

2.1.4 Prozesszeiten

Um Optimierungspotenziale im OP-Umfeld richtig abschätzen und bewerten zu können, ist nicht nur das Verständnis zu den Abläufen und Prozessschritten selbst, sondern auch deren zeitlicher Aufwand relevant. Sowohl Grund- und Regelversorger, Schwerpunktversorger, Maximalversorger als auch Universitätskrankenhäuser führen zu etwa 90 Prozent der Cholezystektomien mit dem laparoskopischen Verfahren durch (vgl. [99]). Aus diesem Grund liegt auch in diesem Abschnitt der Fokus der Analysen auf dieser OP-Art.

Eine eindeutige Aussage zu Zeitangaben bzgl. der gesamten Operationsdauer oder einzelner Prozessschritte kann nur schwer getroffen werden, da die Literatur hierzu große Schwankungen aufweist. So finden sich bspw. in [120] Angaben zur mittleren Schnitt-Naht-Zeit bei der laparoskopischen Cholezystektomie von 48,5 $\pm$ 18,5 Minuten, in [94] werden 55 $\pm$ 17 Minuten genannt und Hunziker et al. geben in [74] sogar 102 $\pm$ 47 Minuten an. Die Gründe für diese Unterschiede können vielfältig sein. Langhorst et al. untersuchen in [99] unter anderem die Varianz der Prozesszeiten von Cholezystektomien in Abhängigkeit der Versorgungsstufe des Krankenhauses. Dort zeigen sie, dass Grund- und Regelversorger in Deutschland insgesamt die meisten konventionell laparoskopischen Cholezystektomien durchführen (11.511 Fälle in der untersuchten Kohorte in den Jahren 2011 - 2013), wobei die durchschnittliche OP-Zeit laut ihrer Analysen bei 59 $\pm$ 27 Minuten liegt. Schwerpunktversorger führen zwar ähnlich viele laparoskopische Cholezystektomien durch (11.386 Fälle), ihre durchschnittliche Schnitt-Naht-Zeit liegt mit 67 $\pm$ 28 Minuten allerdings deutlich höher als die der Grund- und Regelversorger. Obwohl Maximalversorger im Schnitt pro Jahr deutlich weniger Eingriffe durchführen (6.785 Fälle), liegen sie mit einer Durchschnittszeit von 66 $\pm$ 27 Minuten nahe den Schwerpunktversorgern. Universitätskrankenhäuser heben sich mit einer deutlich längeren Schnitt-Naht-Zeit von 89 $\pm$ 39 Minuten bei 1.867 Fällen von den drei Versorgungsstufen ab. Bei der Verteilung des operativen Verfahrens (offen vs. laparoskopisch vs. Umstieg von laparoskopisch zu offen) liegen die Versorgungsstufen nahe zusammen. Lediglich die Universitätskliniken weisen einen wesentlich geringeren Anteil an

laparoskopischen Verfahren von 75,5 Prozent auf. Eine konkrete Begründung für die Varianzen in der Schnitt-Naht-Zeit können Langhorst et al. letztendlich mit den vorliegenden Daten nicht liefern. Als mögliche Gründe nennen sie aber unter anderem die Erfahrung bzw. den Ausbildungsstand des Operationsteams, die Komplexität des vorliegenden Krankheitsbildes oder die Ausstattung des Krankenhauses.

2.1.5 Personalaufwand für die laparoskopische Cholezystektomie

Der Personalaufwand einer laparoskopischen Cholezystektomie setzt sich aus Fachkräften aus verschiedenen Berufsgruppen eines Krankenhauses zusammen. Für den Transport des Patienten von der Station zum OP ist die Stationspflegekraft verantwortlich. In der Schleuse befindet sich zusätzlich eine Pflegekraft, die sich um das Ein- und Ausschleusen des Patienten kümmert. Die Einleitung des Patienten wird sowohl von einem Anästhesisten als auch einer Anästhesiepflegekraft durchgeführt. Im OP-Saal befinden sich zudem weitere Akteure, wie der Chirurg und ein Assistent, ein Anästhesist, ein bis zwei Springer sowie eine instrumentierende Pflegekraft. Der Chirurg und dessen Assistent sind für die Durchführung der Operation verantwortlich und während der gesamten Schnitt-Naht-Zeit anwesend. Der Anästhesist überwacht während der Operation alle Vitalfunktionen des Patienten. Die instrumentierende Pflegekraft verantwortet das Richten der Instrumente auf den Instrumententisch und das Anreichen der sterilen Instrumente. Die zusätzlichen Springer reichen der instrumentierenden Pflegekraft weitere während der Operation benötigte sterile Instrumente und kümmern sich darüber hinaus um alle unsterilen Aufgaben, wie zum Beispiel die Dokumentation des OP-Verlaufes. Bei laparoskopischen Cholezystektomien ist zusätzlich zu den bisher genannten Arbeitskräften außerdem ein OP-Techniker von Bedeutung. Dieser baut den Laparoskopieturm auf und ist während der gesamten OP präsent, falls technische Schwierigkeiten auftreten. Des Weiteren wird Reinigungspersonal benötigt, welches unmittelbar nach Beendigung der OP einsatzbereit ist und den OP-Saal reinigt, desinfiziert und aufräumt.

Grundsätzlich liegen für das unterschiedliche Personal auch verschiedene Präsenzzeiten im Verlauf der laparoskopischen Cholezystektomie vor. Die Anästhesiepflegekraft ist über die Vorbereitung, Einleitung und Lagerung des Patienten, während der Operation und bis hin zur Ausleitung des Patienten anwesend. Der Anästhesist hingegen ist vor allem in der Zeit der Vorbereitung und Einleitung des Patienten und während der Operation präsent. Gegebenenfalls ist er auch zusätzlich bei der Ausleitung des Patienten einsatzbereit. Die Anwesenheit des Chirurgen und der chirurgischen Assistenz ist ausschließlich auf die Zeit der Operation beschränkt. Ihre Arbeit findet somit nur innerhalb des OP-Saals statt. Für die umfangreichen Aufgaben der Springer ist deren Anwesenheit bei der Vorbereitung, Lagerung, Operation und Ausleitung erforderlich, während die instrumentierende Pflegekraft hauptsächlich in der Vorbereitungszeit und bei der Operation anwesend sein muss (vgl. [74]).

2.1.6 Materialien und Instrumente für die laparoskopische Cholezystektomie

2.1.6.1 Laparoskopieturm

Essentiell bei laparoskopischen Operationen ist der Laparoskopieturm, der grundsätzlich bei minimalinvasiven Eingriffen in der Bauchhöhle eingesetzt wird. Auf ihm sind die elektrischen Geräte montiert (vgl. [25]). Dazu zählt ein Hauptmonitor sowie ein Zweitmonitor, ein Videoprozessor mit einer Kaltlichtquelle, der CO_2-Insufflator für das Pneumoperitoneum sowie ein Videodokumentationssystem. Des Weiteren ist eine CO_2-Schlauchverbindung integriert. Dabei besteht der Laparoskopieturm aus einem beweglichen Gestell auf Rollen (vgl. [40]). Da die Kabel und Schläuche der Instrumente von laparoskopischen Eingriffen mit den aufgezählten Geräten des Laparoskopieturms verbunden sind, wird die unmittelbare Nähe des laparoskopischen Operationsturms zum OP-Tisch vorausgesetzt (vgl. [25]).

Der Hauptmonitor am Laparoskopieturm dient hauptsächlich dem Operateur und wird deshalb für diesen stets in gerader und somit ergonomischer Blickrichtung platziert. Ein zweiter Monitor, auf dem ebenfalls die Sicht des Bauchinnenraums abgebildet wird, dient der Assistenzkraft des Operateurs. Dieser Monitor wird meist auf der gegenüberliegenden Seite platziert. Ein zusätzlicher, an der Wand des OP-Saals angebrachter Bildschirm ermöglicht es dem restlichen OP-Personal, den Operationsverlauf genau zu verfolgen (vgl. [40]).

2.1.6.2 Verwendete Instrumente und deren Anordnung am Instrumententisch

Alle für die OP benötigten Instrumente werden auf den dediziert dafür vorgesehenen Instrumententischen angerichtet. Für jede Operation gibt es eine festgelegte, standardisierte Tischaufbauanleitung, nach welcher die instrumentierende Pflegekraft alle Materialien in der vorgegebenen Anordnung für die OP sowie weitere Instrumententische bzw. Beistelltische mit notwendigen Zusatzmaterialien für den Operationsverlauf richtet. Während des Anrichtens werden die die einzelnen Objekte außerdem auf ihre Vollständigkeit und Funktion geprüft. Dabei gilt grundsätzlich, dass der Instrumententisch von links oben nach rechts unten gedeckt wird. Nach dem vollständigen Vorbereiten des Instrumententisches wird dieser im OP-Saal positioniert. die instrumentierende Pflegekraft befindet sich während des gesamten Operationsverlaufes in unmittelbarer Nähe zum Instrumententisch, sodass ein situationsgerechtes Instrumentieren möglich ist. Darüber hinaus ist es zudem die Aufgabe der instrumentierenden Pflegekraft sowohl prä-, intra-, als auch postoperativ eine Zählkontrolle der Instrumente und Materialien durchzuführen (vgl. [49]).

Für den Aufbau des Instrumententisches spielt in erster Linie eine übersichtliche und operationsspezifische Anordnung der Instrumente eine entscheidende Rolle (vgl. [106]). Alle Materialien oder Zusatzinstrumente, die für die Operation nur eventuell benötigt werden, werden auf einem Zusatztisch oder den Beistelltisch angeordnet. Jedes einzelne Instrument, welches bei der Operation verwendet

wurde, wird unmittelbar nach der Nutzung gegebenenfalls oberflächlich gereinigt, bevor es dann an die ursprüngliche Position auf den Instrumententisch zurückgelegt wird. Dies dient einer besseren Übersicht und vereinfacht zudem die Zählkontrolle. Für jede Operation ist festgelegt, dass alle Instrumente, Materialien oder Textilien erst nach der vollständig fertiggestellten Hautnaht den OP-Saal verlassen dürfen und der darauffolgenden Aufbereitung zugeführt werden (vgl. [106]).

Da die instrumentierende Pflegekraft neben dem Richten der Instrumente zusätzlich für das Übergeben und Entgegennehmen der Instrumente verantwortlich ist, muss sie einige Richtlinien berücksichtigen. Um dem Chirurgen ein Instrument zu reichen, muss die Pflegekraft das Instrument stets am Arbeitsteil fassen, sodass direkt die Ringe, der Griff oder der Stil des Instruments gefasst werden können. Nur so kann der Chirurg das Instrument sofort einsetzen. Die instrumentierende Pflegekraft arbeitet dabei beidhändig, so dass sie mit einer Hand dem Chirurgen ein Instrument übergeben und mit der anderen Hand das bereits verwendete und nicht mehr benötigte Instrument entgegennehmen kann (vgl. [106]).

Wie zuvor erwähnt, wird der Instrumententisch von links oben nach rechts unten gedeckt. Bei der laparoskopischen Cholezystektomie werden nach Freese et al. in der oberen Reihe die folgenden Instrumente von links nach rechts angeordnet (vgl. [49]):

- 2 x Haken nach Langenbeck
- 2 x Klemmen nach Mikulicz (groß)
- 1 x Monopolarkabel für die Hakenelektrode
- 1 x Nadelabwurf mit Clips
- 2 x laparoskopische Fasszangen
- 2 x Reduzierhülsen
- 1 x monopolare Schere

- 1 x monopolare Hakenelektrode
- 1 x Clipzange
- 1 x Gallenblasen-Fasszange

In der unteren Reihe werden von links nach rechts die folgenden Instrumente angerichtet (vgl. [49]):

- 1 x monopolarer Elektrodenhandgriff mit Anschlusskabel
- 4 x Trokar
- 1 x Einmalspritze
- 1 x Veress-Kanüle
- 2 x Klemmen nach Backhaus
- 1 x Schere nach Cooper
- 1 x Schere nach Metzenbaum
- 2 x Pinzette
- 1 x Skalpellgriff

Auf dem Beistelltisch befinden sich darüber hinaus das Grundsieb, das Laparoskop, Kompressen, Spreizer, ein Chromaganschälchen mit Tupfern, Pflaster, eine Nierenschale mit Chromaganschälchen für Gallenblase und NaCl 0,9%, ein Nadelsieb, 3 Nadelhalter nach Hegar, Nahtmaterial, ein Lichtkabel und CO_2-Schlauch, Hautdesinfektionsmittel, eine Tupferklemme mit Tupfern sowie das Laparoskopiesieb (vgl. [49]).

Abbildung 2.5 verdeutlicht die Anordnung des OP-Personals, sowie des Instrumententisches, des Beistelltisches und weiterer während einer laparoskopischen Cholezystektomie benötigter Geräte.

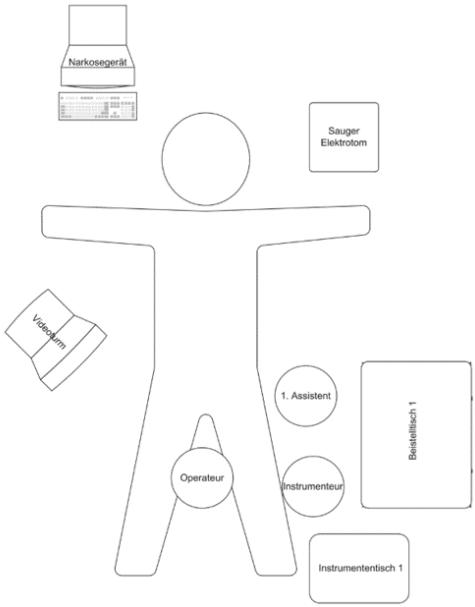


Abbildung 2.5: Saalübersicht einer laparoskopischen Cholezystektomie [49]

2.1.6.3 Instrumentenaufbereitung

Die Instrumentenaufbereitung ist nicht Teil der weiteren Untersuchungen dieser Arbeit, soll aber im Sinne der Vollständigkeit der Abbildung des kompletten OP-Kreislaufs kurz erläutert werden.

Chirurgische Instrumente haben einen hohen monetären Wert und sollen ihre Funktion mittels fachgerechter Aufbereitung so lange wie möglich erhalten. Dabei schreiben regulatorische und gesetzliche Rahmenbedingungen, u. A. die Medical Device Regulation (MDR) [11] und die Medizinproduktebetreiber-Verordnung (MPBetreibV) [1], eine fachgerechte, schriftlich definierte Aufbereitung vor, die Aufbereitungsvorgänge durchweg dokumentiert. Neben den gesetzlichen Regelungen sorgen auch ergänzende Einrichtungen, wie die dem Robert Koch Institut (RKI) zugehörigen Kommission für Krankenhaushygiene

und Infektionsprävention (KRINKO) [91] oder der Nationale Arbeitskreis zur Implementierung der EU-Verordnungen über Medizinprodukte und In-vitro-Diagnostika (NAKI) [15], [113], für Richtlinien, die die Aufbereitungsprozesse sinnvoll und nachhaltig gestalten und gleichzeitig Schäden am Patienten und am OP-Personal vermeiden sollen. Sowohl die Wiederaufbereitung, die Reinigung, die Desinfektion, Pflege, Kontrolle und Sterilisation erfolgen in der zentralen Abteilung für Sterilgutversorgung (ZSVA). Diese Abteilung zeichnet sich durch ihren hohen technischen Standard und ihre speziell qualifizierten Mitarbeitenden aus (vgl. [106]).

Um die Instrumentenaufbereitung zu gewährleisten, durchlaufen die chirurgischen Instrumente den sogenannten Instrumentenkreislauf. Somit werden alle Instrumente nach ihrer Nutzung zuerst zerlegt, bevor sie gereinigt und anschließend desinfiziert werden. Die damit verbundene Pflege der Instrumente umfasst außerdem eine Funktionsprüfung, bevor der Prozess der Sterilisation beginnt. Auch dabei ist eine sorgfältige Dokumentation notwendig. Erst im Anschluss ist die Lagerung der Instrumente für eine Operation und somit die Bereitstellung für deren weitere Nutzung möglich. Der Aufbereitungsprozess wird dabei maschinell unterstützt (vgl. [106]).

2.1.7 Das Krankenhausinformationssystem

Wesentlicher Bestandteil zum Datenaustausch innerhalb der Krankenhaus-IT ist das Krankenhausinformationssystem. Dabei handelt es sich um ein „computergestütztes Primärsystem für abteilungsübergreifende Erfassung, Weiterbearbeitung und Archivierung von Informationen [. . .] innerhalb des Krankenhauses“ [79]. Mit Hilfe des KIS sollen administrative und klinische Arbeitsprozesse im Krankenhaus optimiert werden. Durch die Integration von Informationen in einer Datenbank der medizinischen, pflegerischen und administrativen Bereiche des Krankenhauses gelingt ein computerbasierter Datenaustausch mit anderen Systemen. Das KIS selbst beinhaltet ein *Klinisches Arbeitsplatzsystem (KAS)* für

den Informationszugriff durch das Krankenhauspersonal, das *Patientendatenmanagementsystem (PDMS)* mit Patientendatenbank zur Patientenadministration, das *Picture Archiving and Communication System (PACS)* zur Verwaltung aller medizinischen Bilddaten sowie ein digitales Archivierungssystem als Langzeitdatenspeicher und einen zentralen Kommunikationsserver. Letzterer unterstützt den Datenaustausch zwischen den verschiedenen Teilkomponenten des KIS [79]. Für die Kommunikation zwischen den einzelnen Teilsystemen kommen üblicherweise kommerzielle Standardschnittstellen wie *Health Level 7 (HL7)* oder speziell für Bilddaten *Digital Imaging and Communications in Medicine (DICOM)* zum Einsatz. Die dargelegten Zusammenhänge sind in Abbildung 4.12 skizziert.

2.2 Kontexterfassung

Ergänzend zu den bisher diskutierten medizinischen Grundlagen werden im Folgenden die notwendigen informationstechnischen Voraussetzungen aufbereitet. Zum besseren Verständnis der nachfolgend erläuterten Konzepte sollen zunächst einige grundlegende Begriffe zur Kontexterfassung für diese Arbeit definiert werden. Des Weiteren gibt dieses Kapitel einen Überblick über häufig eingesetzte Sensorik sowie die gängigsten Methoden, um bestimmte Kontexte automatisiert zu erkennen und analysieren zu können.

2.2.1 Definition Kontext & Kontextsensitivität

Unter *Kontext* werden alle Informationen verstanden, die eine konkrete Situation beschreiben. Beispiele für solche Informationen sind u. A. Lokation, Aktivität, Handlungen und Prozesse, Interaktion mit/zwischen Objekten, aber auch Vitalparameter oder Sprache und Geräusche.

Kontextsensitivität im informationstechnischen Sinne beschreibt die Fähigkeit eines technischen Systems der Adaption auf den aktuellen Kontext und somit

der Reaktion auf Veränderungen der Umwelt [137]. Diese Definition wird auch von Abowd et al. in [2] gestützt:

> „A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user‘s task.“

Folglich beschäftigt sich die *Kontexterfassung* allgemein beschrieben mit dem automatisierten Erfassen der für den jeweiligen Kontext relevanten Informationen basierend auf Beobachtung. Die Kontexterfassung ist ein sehr breites Feld und stark abhängig von den jeweiligen Umgebungsbedingungen. Zur weiteren Verdeutlichung werden in den folgenden Abschnitten noch Details zu möglicher Sensorik und relevanten Methoden diskutiert. Darüber hinaus können auch Informationen aus nicht-sensorischer Erfassung einbezogen werden. Im medizinischen Umfeld können dies insb. Angaben aus dem KIS aber auch anderen Dokumentationssystemen sein.

2.2.2 Sensorik zur Kontexterfassung

In diesem Abschnitt werden verschiedene Kategorien von Sensorik zur Kontexterfassung erläutert. Grundsätzlich können in diesem Umfeld zwei Gruppen von Sensoren unterschieden werden:

1. **Körpernahe Sensorik:** Sensoren, die direkt am Körper bzw. an einem Objekt angebracht werden und somit auf Aktionen des Trägers reagieren
2. **Ambiente Sensorik:** Sensoren, die in der Umwelt der Person angebracht sind und ihr Umfeld beobachten.

Abbildung 2.6 skizziert diese Einteilung und nennt jeweils Beispiele konkreter Instanzen. Dabei wird auch deutlich, dass diese, je nach Zielsetzung und Anwendungsfeld, beiden Gruppen angehören können. Die genannten Sensoren bilden dabei nur eine Auswahl und zeigen häufig genannte Beispiele. Darüber hinaus können auch andere Sensoren zur Kontexterkennung eingesetzt werden.

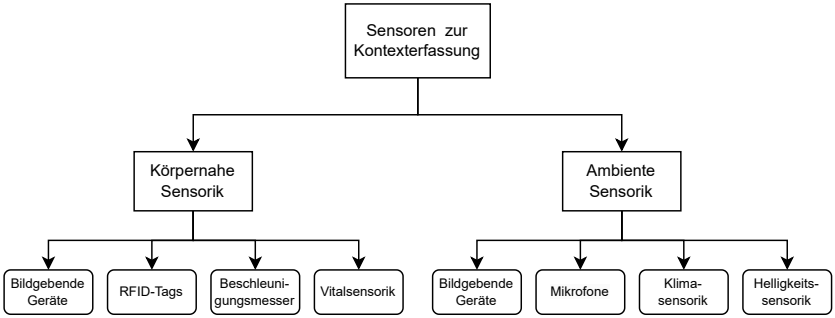


Abbildung 2.6: Einteilung von Sensoren zur Kontexterkennung, angelehnt an [97]

Üblicherweise sind die körpernahen Sensorinformationen konkreter der relevanten Information zuzuordnen und entsprechend einfacher für die Datenauswertung nutzbar. Jedoch sind sie in der Handhabung eingeschränkter. So muss eine adäquate und, bei beweglichen Objekten, kabellose Stromversorgung sichergestellt sein. Dies wiederum erfordert regelmäßiges Laden der Akkus bzw. Austauschen der Batterien. Des Weiteren muss die Anbringung gleichzeitig stabil, möglichst unaufdringlich und nicht störend oder einschränkend sein. Bei nicht permanenter Befestigung spielt zusätzlich noch die Möglichkeit eine Rolle, dass der Sensor, absichtlich oder unabsichtlich, nicht getragen wird oder auch verloren gehen kann.

Im Gegensatz dazu sind ambiente Sensoren meist unaufdringlicher und weniger komplex in der Benutzung, die gewonnenen Daten dabei aber oftmals abstrakter, sodass die Sensorwerte nicht direkt dem gewünschten Kontext zugeordnet werden können.

Aus diesen Erläuterungen wird deutlich, dass die Auswahl der Sensorik für ein Kontexterfassungssystem stark abhängig von vielen verschiedenen Faktoren, wie z. B. der vorherrschenden Umgebung, der Zielgruppe oder auch der verfügbaren Ressourcen, ist. Dabei können auch mehrere Einzelsensoren zu einem komplexen Sensornetzwerk zusammengeführt werden, z. B. um den Erfassungsbereich zu vergrößern oder um den Informationsgehalt der Daten zu erhöhen.

Im Verlauf der vorliegenden Arbeit werden als Informationsquelle hauptsächlich Kameras in verschiedenen Varianten zur Aufzeichnung von Video- und Bilddaten eingesetzt. Daten aus bildgebenden Sensoren haben, bei sorgfältig ausgewählter Auslegung, einen extrem hohen Informationsgehalt. Die Herausforderung besteht darin, diese Informationen automatisiert auszuwerten. Das *maschinelle Sehen* (engl. *Computer Vision*) durch Bild- und Videoanalyse ist ein Forschungsfeld, bei welchem noch nicht alle Probleme effizient gelöst sind. Gleichzeitig stehen viele Menschen einer visuellen Überwachung skeptisch gegenüber, da bei dieser Technologie sowohl Datenschutz- als auch Datensicherheitsaspekte häufig besonders sensibel hinterfragt werden. Auf eine Beschreibung der technischen und optischen Grundlagen von Kameras wird an dieser Stelle verzichtet und bspw. auf [63] und [67] verwiesen.

2.2.3 Methoden der Kontexterfassung

Methodisch können zwei mögliche Ansätze zur Datenanalyse unterschieden werden:

1. Wissensgetriebene Analyse
2. Datengetriebene Analyse

Wissensgetriebene Systeme werden häufig auch als *Expertensysteme* bezeichnet, da für ihre Realisierung eine große Menge an Expertenwissen aus der Anwendungsdomäne gesammelt und ausgewertet wird [43]. Basierend darauf werden Merkmale händisch extrahiert, die ein Modell zur Abbildung der Messwerte auf die Realität definieren. Der Vorteil von Expertensystemen ist, dass sie deterministisch verlaufen und der Auswerteprozess somit vollumfänglich nachvollzogen werden kann. Dafür ist zur Gewinnung und Verarbeitung des Expertenwissens der Zeitbedarf sehr hoch und auch abhängig vom Zugang zu den benötigten Informationen, bspw. durch enge Zusammenarbeit mit Experten aus der Anwendungsdomäne.

Bei datengetriebenen Methoden wird häufig auch von maschinellen Lernverfahren (engl. *machine learning*) gesprochen. Diese wiederum lassen sich in die Unterkategorien überwachtes (engl. *supervised*), unüberwachtes (engl. *unsupervised*) und bestärkendes Lernen (engl. *reinforcement learning*) gliedern [39]. Maßgeblich beim maschinellen Lernen ist, dass die Merkmale, welche das gewünschte Modell beschreiben, automatisiert aus Beispieldaten extrahiert werden. In der Lernphase werden dafür Muster in den Beispieldaten erkannt, die später in der Inferenzphase auch auf unbekannte Daten angewendet werden, wodurch die gesuchten Informationen extrahiert werden können. Für eine genauere Beschreibung dieser Verfahren sei u. A. auf [42], [60] oder [116] verwiesen. Im Gegensatz zu den wissensgetriebenen Modellen sinkt bei diesem Ansatz der Bedarf an Expertenwissen erheblich. Außerdem können sehr viel komplexere und genauere Modelle gewonnen werden. Allerdings sind beim maschinellen Lernen die Anforderungen sowohl an die Qualität und Quantität der Beispieldaten als auch an die Rechenleistung zum Trainieren der Modelle sehr hoch. Zudem sind die aus maschinellen Lernverfahren gewonnenen Modelle in vielen Fällen nicht deterministisch und entsprechend nur schwer verständlich. Dies führt vor allem in regulierten Umfeldern, wie der Medizintechnik oder der Automobilbranche, zu zusätzlichen Herausforderungen.

Somit ist, neben der Auswahl der Sensorik, auch die Wahl der Methodik stark abhängig von verschiedenen Faktoren, wie bspw. der Daten- und Wissenslage oder den Hardwareressourcen.

Aufgrund der Forschungserfolge der letzten Jahre auf dem Gebiet des maschinellen Lernens und insb. des *Deep Learning (DL)* generell bei der Datenanalyse und auch beim maschinellen Sehen im Speziellen, liegt im weiteren Verlauf dieser Arbeit der Fokus auf eben jenen Methoden. In den folgenden Unterabschnitten werden die hier relevantesten Grundlagen kurz skizziert. Die Auswahl orientiert sich an den benötigten Informationen bzgl. der Kontexte der einzelnen Teilsysteme dieser Arbeit (vgl. Kapitel 4).

2.2.3.1 Objekterkennung

Eine der bekanntesten Anwendungen des maschinellen Sehens ist die Objekterkennung. Ikeuchi definiert dies in [75] wie folgt:

> „Bei der Objekterkennung geht es darum, Instanzen von Objekten einer bestimmten Klasse in einem Bild zu erkennen."

In den 2000er Jahren nutzten gängige Lösungen für die Objekterkennung Merkmalsdeskriptoren wie die 1999 von Lowe entwickelte Scale-Invariant Feature Transform (SIFT) [108] und das 2005 populär gewordene Histogram of Oriented Gradients (HOG) [31]. In den 2010er Jahren fand dann eine Verlagerung hin zur Verwendung von Convolutional Neural Networks (CNNs) statt [96], [141], [147]. In [60] werden Faltungsnetzwerke (dt. für CNNs) als neuronale Netze bezeichnet, die in mindestens einer ihrer Schichten die Faltung anstelle der allgemeinen Matrixmultiplikation verwenden. Faltungen sind lineare Operationen, die besonders gut für Daten mit einer gitterartigen Struktur, wie bspw. Bilder, geeignet sind. Der Ablauf der Faltung wird an dieser Stelle nicht explizit erläutert und kann der einschlägigen Literatur entnommen werden (z. B. [60]). Abbildung 2.7 stellt schematisch den gesamten Ablauf einer CNN-basierten Objekterkennung dar.

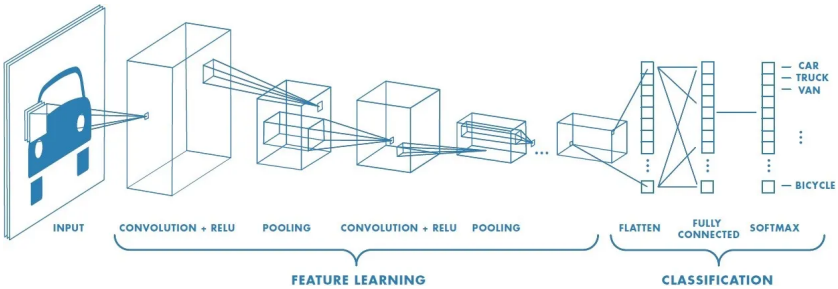


Abbildung 2.7: Schematische Darstellung einer Objektklassifizierung mittels CNN [149]

2.2.3.2 Personen- und Posenerkennung

Da zur Analyse von OP-Abläufen u. A. auch die Handlungen von und Interaktionen zwischen den Akteuren eine hohe Relevanz aufweisen, ist das Erkennen von Personen und deren Körperhaltung und Posen ein wichtiger Bestandteil. Zum reinen Erkennen von Personen in Bildern und Videos sind die zuvor genannten Methoden der Objekterkennung bereits ausreichend und in der Forschung weit fortgeschritten. Sollen darüber hinaus noch die Körperhaltung bzw. die Position und Orientierung einzelner Körperteile im Raum erfasst werden, kommen üblicherweise Methoden der Posenerkennung (engl. *Pose Detection*, in der Fachliteratur oft auch *Pose Estimation*) zum Einsatz. Das Ergebnis der Posenerkennung wird häufig in Form einer Skelettstruktur dargestellt, die sich aus der Verbindung der erkannten *Keypoints*, die einzelne Körperteile repräsentieren, ergibt. Abbildung 2.8 zeigt eine solche Darstellung. Hier wird auch die Nummerierung der Keypoints nach dem *COCO Keypoint Format* ersichtlich.

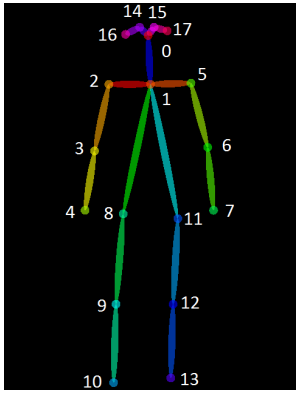
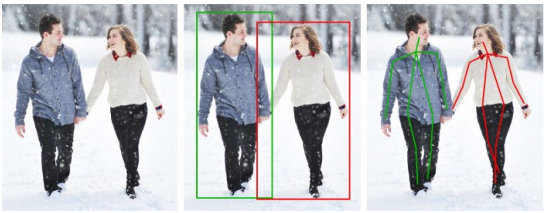


Abbildung 2.8: Beispielhafte Darstellung einer mittels Posenerkennung gewonnenen Skelettstruktur eines Menschen im COCO Keypoint Format (Quelle: [28]).

Für die Posenerkennung sind in der Literatur zwei Herangehensweisen zu finden [124]:

1. **Top-Down:** Hierbei werden zunächst die Personen als Ganzes detektiert. Anschließend werden in den jeweiligen Bildausschnitten die unterschiedlichen Keypoints lokalisiert, um daraus das Skelett der Person zu erstellen (vgl. Abbildung 2.9a (oben)).

2. **Bottom-Up:** Bei dieser Herangehensweise werden zuerst alle Keypoints im Bild detektiert. Anschließend werden diese gruppiert, um daraus für jede Person ein Skelett zu erstellen (vgl. Abbildung 2.9b (unten)).



(a) Top-Down

(b) Bottom-Up

Abbildung 2.9: Herangehensweise Pose Estimation: Top-Down (oben) und Bottom-Up (unten) (Quelle: [124])

Als Erweiterung der reinen Posenerkennung kann das Posentracking (engl. *Pose Tracking*) betrachtet werden. Es beschreibt allgemein die Aufgabe, die Körperhaltung von Personen in Videos abzuschätzen und Personen über mehrere Einzelbilder hinweg zu identifizieren sowie zu unterscheiden. Dafür wird jeder Person eine individuelle Identifikationsnummer (ID) zugeordnet mit welcher die Bewegungen über einen längeren Zeitraum hinweg im Video verfolgt werden können. Bei erfolgreichem Tracking behält jede Person ihre ID, so lange, bis sie den Bildausschnitt verlässt. Neu hinzukommende Personen erhalten

jeweils eine neue ID. Dies gilt im Regelfall auch für Personen, die den Bildausschnitt verlassen haben und später erneut sichtbar werden. Abbildung 2.10 zeigt beispielhaft ein Einzelbild eines Videos, auf das Posentracking für mehrere Personen gleichzeitig (sog. *Multi-Person Pose Tracking*) angewendet wurde. Es ist zu erkennen, dass jeder Person eine eigene ID zugewiesen wurde. Durch das Tracking ist es möglich, Bewegungen einzelner Personen zu extrahieren. Posentracking liefert eine genaue Abschätzung menschlicher Bewegungen und dient beispielsweise der Erkennung menschlicher Handlungen, dem Verständnis zwischenmenschlicher Interaktion sowie der Bewegungsdetektion [117].

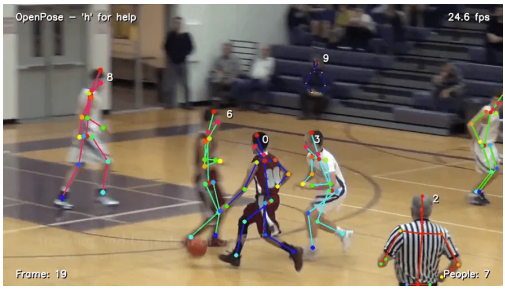


Abbildung 2.10: Beispielbild für Multi-Person Pose Tracking (Quelle: [73])

Die meisten Verfahren zur Pose Estimation oder zum Pose Tracking basieren, ähnlich wie bei der Objekterkennung, grundlegend auf der Verwendung von CNNs.

2.2.3.3 Re-Identifikation von Personen

Wie im vorherigen Abschnitt beschrieben, bieten klassische Methoden für Personenerkennung oder -tracking keine Möglichkeit, einzelne Personen nach Verlassen und Wiedereintreten in den Bildausschnitt wieder derselben ID zuzuordnen. Dies liegt daran, dass die spezifischen Merkmale von Individuen bei diesen Verfahren nicht berücksichtigt werden. Mit dieser Herausforderung

beschäftigen sich Methoden zur *Re-Identifikation* von Personen, die in diesem Abschnitt detaillierter betrachtet werden.

Die Re-Identifikation (Re-ID) von Personen (engl. *Person Re-Identification*) bezeichnet im Bereich der Computer Vision die Aufgabe, Personen anhand optischer Merkmale in Bild- oder Videoaufnahmen zu identifizieren, nachdem diese zuvor bereits erfasst wurden. Optische Merkmale zur Re-Identifikation einer Person können dabei sowohl aus Bild- und Videodaten als auch aus Beschreibungen in Textform entnommen werden [166]. In der Literatur finden sich unterschiedliche Definitionen des Anwendungsfalls. Die Re-Identifikation von Personen wird dabei entweder in einem System aus nicht überlappenden Kameras durchgeführt [59], [166] oder mit einer einzelnen Kamera, wobei die Personen mehrmals zu unterschiedlichen Zeitpunkten zu sehen sind [166].

Das Ablaufdiagramm in Abbildung 2.11 stellt die allgemeine Funktionsweise eines Re-Identifikationssystems dar. Die nachfolgende Beschreibung bezieht sich auf Systeme, welche basierend auf Bild- oder Videodaten arbeiten.

Ziel der Re-Identifikation von Personen ist es, eine Testperson anhand einer Bildaufnahme (das sog. *Query*) in einer Datenbank an Aufnahmen (der sog. *Galerie*) zu identifizieren. Die Aufnahme des Queries ist dabei üblicherweise nicht identisch mit der Aufnahme aus der Galerie, sondern unterscheidet sich in verschiedenen Eigenschaften, wie Perspektive, Beleuchtung, Haltung und Tätigkeit der Person.

Das System erhält zu Beginn Bild- oder Videodaten einer Kamera. Für die weitere Verarbeitung ist es notwendig, einzelne Personen in den Bilddaten zu detektieren. Dies kann durch den Einsatz von Algorithmen zur Personendetektion erreicht werden (vgl. Abschnitte 2.2.3.1 & 2.2.3.2). Der Detektionsprozess ist im Allgemeinen nicht Teil der Re-ID und wird in der Literatur häufig als vorausgesetzt betrachtet [59]. Abbildung 2.12 zeigt beispielhaft die Extraktion von Einzelbildern aus einer Gesamtaufnahme.

Die extrahierten Einzelbilder bilden den Ausgangspunkt für die Re-ID der Personen. Der erste Schritt ist die Extraktion unterscheidbarer Merkmale für

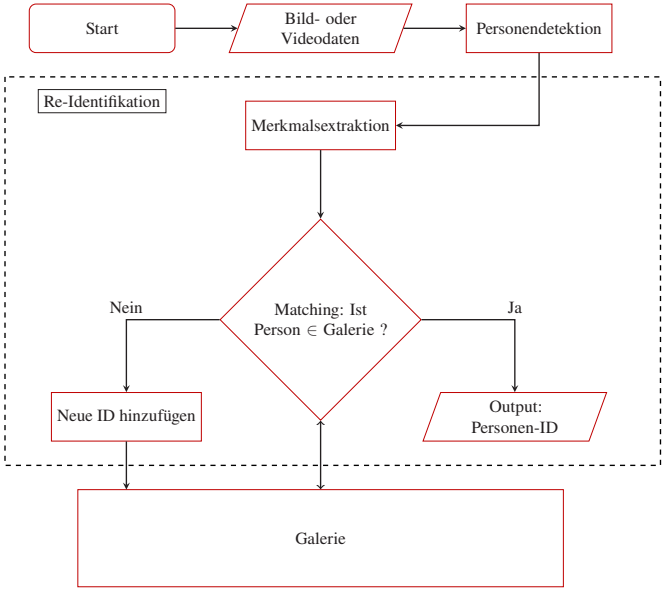
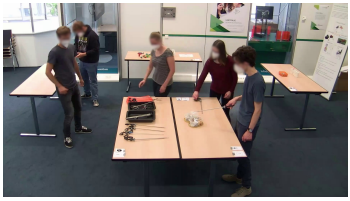


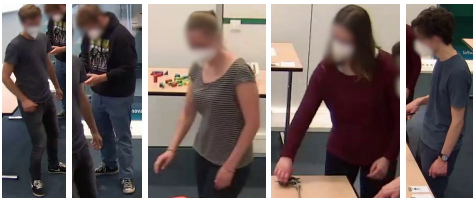
Abbildung 2.11: Aufbau eines Re-Identifikationssystems. Der schwarze Rahmen markiert den Re-Identifikationsprozess. Angelehnt an [9].

verschiedene Personen. Merkmale, die aus den Einzelbildern extrahiert werden können, beschreiben vor allem die Farb- und Textureigenschaften des Bildes. Die Merkmalserzeugung kann dabei bewusst beeinflusst werden, indem verschiedene statistische und mathematische Methoden eingesetzt werden. Einige dieser Methoden werden in Kapitel 3 genauer vorgestellt. Durch den Einsatz von Deep-Learning-Ansätzen lassen sich hochunterscheidbare Merkmale erzeugen, welche den „von Hand“ erzeugten Merkmalen überlegen sind [170]. Die Funktionsweise dieser Algorithmen ist in Abschnitt 3.2.2 genauer beschrieben.

Die erzeugten Merkmale werden im nächsten Schritt für das sog. *Matching* verwendet. Während des Matchings werden die Merkmale des Queries mit den Merkmalen der Personen aus der Galerie verglichen. Für jede Person in der Galerie ist eine individuelle ID hinterlegt. Ziel des Matchingprozesses ist es, das Query in den Aufnahmen der Galerie wiederzuerkennen. Dadurch kann dem Query eine ID zugeordnet werden. Für den Zuordnungsprozess werden von



(a) Gesamtaufnahme



(b) Einzelbilder

Abbildung 2.12: Mithilfe von Algorithmen zur Personendetektion werden aus der Gesamtaufnahme (2.12a) Bildausschnitte der einzelnen Personen erzeugt (2.12b). Die Einzelbilder dienen als Input für das Re-Identifikationssystem.

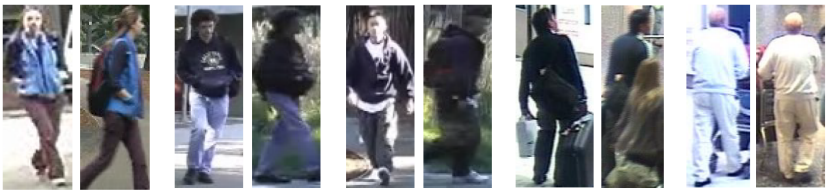
jeder Person in der Galerie ebenfalls Merkmale extrahiert und anschließend mit den Merkmalen des Queries verglichen. Die genaue Funktionsweise des Matchingprozesses wird in Abschnitt 3.2.2 beschrieben.

Ist das Matching erfolgreich, wird dem Query die ID zugeordnet, die für die entsprechende Person in der Galerie hinterlegt ist. Kann das Query nicht zugeordnet werden, wird es je nach Ausführung des Systems unter einer neuen ID der Galerie hinzugefügt oder verworfen.

Bild- und Videoanalysen und die Re-Identifikationsaufgabe im Speziellen bergen charakteristische Herausforderungen, welche durch das Re-ID-System gelöst werden müssen.

Häufig auftretende Probleme, welche die Re-ID erschweren, sind:

- **Inter-Klassen-Variation:** Verschiedene Personen besitzen/erzeugen in einer Aufnahme ein ähnliches oder dasselbe Erscheinungsbild.



(a) (b) (c) (d) (e)

Abbildung 2.13: Herausforderungen bei der Re-Identifikation von Personen. Jeweils ein Bildpaar (a - e) zeigt zweimal dieselbe Person. Zu sehen sind Haltungs- und Perspektivvariationen in (a), (b) und (c), sowie teilweise Verdeckung (d) und Beleuchtungs- und Farbunterschiede (e). [136]

- **Intra-Klassen-Variation:** Dieselbe Person besitzt/erzeugt in verschiedenen Aufnahmen unterschiedliche Erscheinungsbilder.
- **Verdeckung:** Personen werden zum Großteil oder teilweise von Gegenständen oder anderen Personen verdeckt.
- **Variation der Bildeigenschaften:** Eigenschaften wie Beleuchtung oder Farbwiedergabe sind abhängig von der verwendeten Kamera sowie von Umgebungsbedingungen (Wetter, Lichtquelle etc.) und können sich mit der Zeit ändern.

Beispiele für die verschiedenen Ausprägungen sind in Abbildung 2.13 dargestellt.

Des Weiteren werden an Re-Identifikationssysteme Anforderungen bezüglich der Generalisierbarkeit für unterschiedliche Anwendungsbereiche sowie an die Skalierbarkeit für große Datenmengen gestellt [59], [136].

2.2.3.4 Aktivitätserkennung

Ergänzend zur Erkennung von Objekten und Personen bildet die Erfassung von Aktivität und konkreter Handlungen einen zentralen Schritt für das automatische Verständnis von komplexen Szenen im beobachteten Umfeld. In der Literatur ist dieser Vorgang als *Aktivitätserkennung* oder *Handlungserkennung* (engl. *Activity*

Recognition, *Action Recognition* oder, speziell bei Fokus auf menschliche Interaktion, auch *Human Action Recognition (HAR)*) bekannt. Ikeuchi verwendet in [75] folgende Definition:

> „Unter Aktivitätserkennung wird der Prozess der Identifizierung der von Menschen über einen bestimmten Zeitraum ausgeführten Bewegungstypen verstanden. Sie wird auch als Handlungserkennung bezeichnet, wenn die Zeitspanne relativ kurz ist."

Dabei ist das Ziel die Prädiktion einer Klasse einer Aktion eines Individuums oder einer Gruppe von Individuen in Videos. Die Tatsache, dass sich Menschen bei derselben Aktion unterschiedlich verhalten, die Aktion aus unterschiedlichen Perspektiven geschehen kann und Aktionen ähnlich sein können, stellt eine besondere Herausforderung dieser Aufgabe dar [51]. Ein wichtiger Aspekt der Aktivitätserkennung ist die zeitliche Abhängigkeit. Ein möglicher Ansatz, diese abzubilden, ist die Nutzung von 3D-Faltungskernen zusätzlich zu den 2D-Varianten innerhalb der CNNs. Die dritte Dimension wird dabei durch mehrere konsekutive Frames eines Videos gebildet, was jedoch auch zu erheblich höherem Rechenaufwand führt. Eine weitere Möglichkeit zur Repräsentation temporaler Merkmale ist der *optische Fluss* (engl. *Optical Flow*), welcher die Veränderungen im Bild zwischen aufeinanderfolgenden Frames auf Pixelebene berechnet. In der Praxis werden oftmals beide Ansätze kombiniert.

Um die Aktionserkennung mittels CNNs umzusetzen, wurden von Karpathy et al. mehrere Methoden vorgestellt, die räumliche temporale Informationen von konsekutiven Frames zusammenfassen. Die Architekturen der in 2.14 abgebildeten Methoden zeigen unterschiedliche Möglichkeiten zur Erfassung von räumlichen temporalen Features. Dabei stellt die unterste Schicht die einzelnen Frames eines Videos als Eingabe und die letzten Schichten die Klassifikationsschicht als Ausgabe dar. Bei der ersten Methode, der *Single Frame Fusion*, werden alle Informationen der Frames in der letzten Schicht zusammengefasst. Die *Early Fusion* stellt eine Erweiterung der Single Frame Fusion dar, indem mehrere Frames bei der Eingabe zusammengefasst werden. Bei der *Late Fusion* hingegen werden zwei Netze mit geteilten Gewichten verwendet, wobei jedes

Netz unterschiedliche Eingabedaten erhält und diese in der letzten Schicht zusammenfasst. In der letzten Methode, der *Slow Fusion*, werden mehrere Frames durch mehrere Schritte immer weiter zusammengefasst [88].

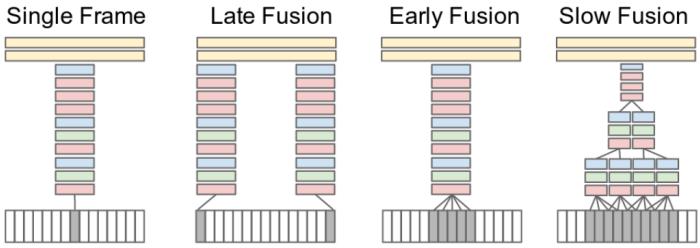


Abbildung 2.14: Zusammenfassen von Informationen über die Zeit. Rote, grüne und blaue Boxen entsprechen den Convolution, Normalisierung und Pooling Schichten nach Karpathy et al. [88]

Aufbauend auf den Erkenntnissen von Karpathy et al. entwickelten Simonyan and Zisserman ein Modell, das auf mehreren parallelen Erkennungs-Streams basiert. Konkret haben sie ein Modell für die Erkennung räumlicher Kontexte und ein weiteres für die Erkennung von Bewegungen auf Basis des optischen Flusses separat trainiert und anschließend mittels Support Vector Machine (SVM) fusioniert (vgl. Abbildung 2.15) [140].

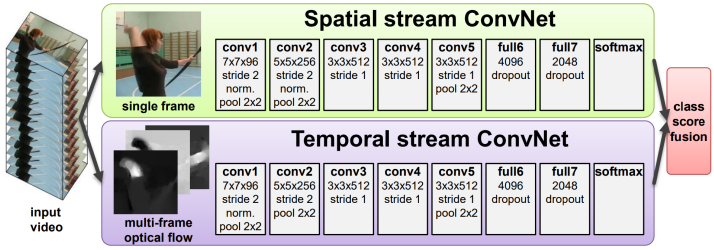


Abbildung 2.15: Zwei-Stream-Architektur für Video-Klassifikation [140]

Auf diesen beiden Konzepten bauen zahlreiche zum Zeitpunkt der Erstellung dieser Ausarbeitung relevanten Arbeiten auf dem Gebiet der Handlungserkennung auf.

Ergänzend zu den spezifischen Herausforderungen der Re-ID birgt die Aktivitätserkennung laut [92] allgemeiner betrachtet außerdem folgende spezifische Herausforderungen:

- **Hintergrundrauschen und Kamerabewegung:** Nicht einheitliche Bildhintergründe und Bewegungen der Kamera beeinflussen die Aktivitätsmerkmale und erschweren dadurch die Erkennung.
- **Unzureichende annotierte Daten:** Zur Generalisierung von gut funktionierenden Modellen in Laborumgebungen hin zu realen Anwendungen sind sehr viel größere Datenmengen notwendig als oftmals verfügbar.
- **Aktions-Vokabular:** Aktionen und Aktivitäten können auf verschiedene Art und Weise definiert werden. Zum Modelltraining sind einheitliche und eindeutige Beschreibungen notwendig.
- **Ungleiche Berechenbarkeit:** Einerseits haben in Videos nicht alle Frames gleichermaßen große Aussagekraft bzgl. einer Aktivität. Viele Frames sind redundant in ihrem Informationsgehalt. Andererseits sind unterschiedliche Aktivitäten u. U. auch unterschiedlich gut erkennbar.

Diese Herausforderungen erschweren die Entwicklung robuster Methoden zur Aktivitätserkennung und müssen stets beachtet werden.

2.2.3.5 Sequenzanalyse

Für die Analyse von Sequenzen, im vorliegenden Anwendungsfall bspw. von einzelnen Handlungen oder eingesetzter OP-Instrumente, ist es notwendig temporale Abhängigkeiten, also vorangegangene Informationen, in die Problemlösung einzubeziehen. Um dies mit neuronalen Netzen zu lösen, wurden schon in den

1980er Jahren Architekturen präsentiert, welche mit sequentiellen Daten umgehen können und deren temporalen Kontext erfassen [86]. Dabei sind die Ziele dieser Verfahren die Vorhersage des nächsten Elementes, die Klassifizierung oder das Generieren gänzlich neuer Sequenzen [64].

Eine Möglichkeit der Modellierung temporaler Zusammenhänge sind sog. *Rekurrente neuronale Netzwerke (RNNs)*. Diese ermöglichen durch schleifenartige (*rekurrente*) Verbindungen die Nutzung von Ausgaben bestimmter Schichten des Netzwerks als Eingaben nachfolgender Netzwerkdurchläufe und somit die Persistenz vergangener Informationen. Ein bekanntes Problem von RNNs ist jedoch das *Vanishing Gradient Problem*, wodurch der Gradient bei der Backprobagation im Training bei langen temporalen Abhängigkeiten verschwindend klein wird und dadurch die Gewichte des Netzes nicht mehr ausreichend beeinflusst, um weitere Trainingsfortschritte erzielen zu können [69]. Nach Hochreiter sind RNNs demzufolge nicht fähig Probleme zu lösen, welche eine lange temporale Abhängigkeit aufweisen. Abbildung 2.16 illustriert den Aufbau eines RNNs.

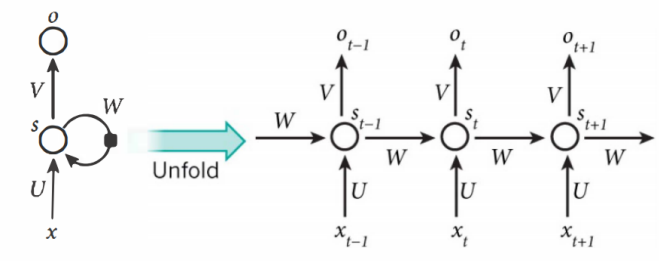


Abbildung 2.16: Struktur rekurrenter neuronaler Netze (Quelle: [101])

Als Lösung für das Vanishing Gradient Problem stellen Hochreiter and Schmidhuber das *Long-Short-Term-Memory Netz (LSTM)* vor, das eine erweiterte RNN-Architektur ist und lange temporale Abhängigkeiten erfassen kann. Um den konstanten Fehlerfluss, welcher durch das Vanishing Gradient Problem verhindert wird, zu gewährleisten, wird das *Constant Error Carrousel (CEC)* eingeführt, indem zusätzliche Features zur RNN-Architektur hinzugefügt werden. Dabei werden Neuronen zu einer Speicherzelle, auch LSTM-Zelle genannt,

gruppiert, wobei der Fehlerfluss durch sog. *Gates* konstant gehalten wird [71]. Die einzelnen Gates kontrollieren dabei den Fluss der Informationen der vorherigen und aktuellen Eingaben und Ausgaben. Somit wird eine einzelne rekurrente Einheit durch eine LSTM-Zelle mit drei Gates und einem Zellenstatus ersetzt [54], [71] (vgl. Abbildung 2.17). Durch die Erweiterung der einfachen rekurrenten Einheit eines RNNs zur LSTM-Zelle können mittels der Gatestruktur und dem Speichern des aktuellen Kontextes, dem Zellzustand, Langzeitabhängigkeiten modelliert werden, ohne dass dabei das Vanishing Gradient Problem auftritt [70]. Zusätzlich ist im Zusammenhang mit LSTMs zu erwähnen, dass vereinfachte Formen der LSTM-Gatestruktur existieren. Die bekannteste Variante stellen hierbei die *Gated Reccurent Units (GRUs)* dar, bei denen die Anzahl der Gates und Parameter im Vergleich zum LSTM zur Beschleunigung des Trainings reduziert werden [27].

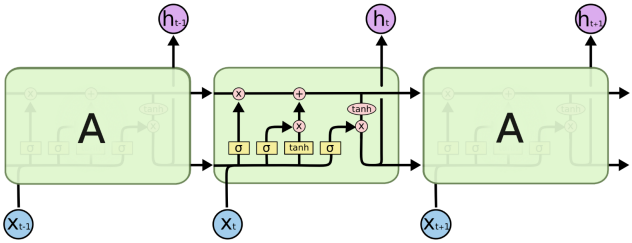


Abbildung 2.17: Aufbau und Zusammenhang mehrerer LSTM-Zellen (Quelle: [118])

Speziell in der Sequenzmodellierung, also der Erzeugung einer neuen Sequenz aus einer vorhandenen Sequenz, wird, abgrenzend zu den rekurrenten Netzen, häufig das Temporal Convolutional Network (TCN), eine vereinfachte Form des WaveNets [33] aus dem Text-to-Speech-Umfeld, eingesetzt [7], [100]. TCNs unterscheiden sich zu gewöhnlichen CNNs sowohl in der Funktionsweise als auch der Architektur. Zum einen sind die Faltungen der TCNs kausal, wodurch keine Informationen aus der Zukunft in die Vergangenheit geraten können und zum anderen kann als Eingabe jegliche Sequenzlänge verwendet und diese einer Ausgabe derselben Länge zugeordnet werden [7]. Durch die Kausalität ergeben sich jedoch Einschränkungen für das TCN, wodurch Aufgaben wie maschinelles

Übersetzen oder Sequence-to-Sequence Vorhersagen nicht abgebildet werden können, da zukünftige Eingaben für diese Art von Aufgaben eine Rolle spielen. Dennoch kann dies nach Bai et al. durch Anpassungen der Architektur ermöglicht werden. Um temporale Abhängigkeiten für das Netzwerk erfassbar zu machen, wird anstatt der normalen Faltung eines CNNs die *Causal Dilated Convolution* des WaveNets verwendet [7]. Mit steigender Anzahl *Dilatationen* steigt dabei die Größe des rezeptiven Feldes, was dem Netz ermöglicht möglichst viele Daten aus der Vergangenheit zu erfassen [33]. In Abbildung 2.18 ist ein solches TCN dargestellt.

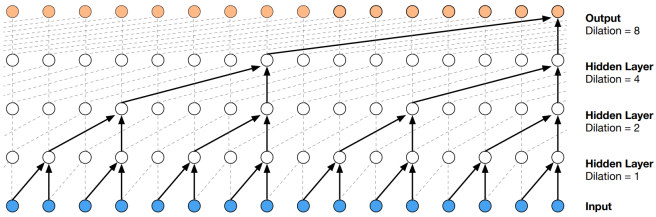


Abbildung 2.18: Schematischer Aufbau eines TCNs (Quelle: [33])

Im Bereich der Verarbeitung natürlicher Sprache (*Natural Language Processing (NLP)*), einer beliebten Anwendung von Sequenzanalyse, haben sich auf Self-Attention basierende Architekturen, insbesondere *Transformer*, aufgrund ihrer Rechenleistung und Skalierbarkeit als bevorzugte Modelle herausgestellt. Anstelle von rekurrenten Einheiten und Faltungen nutzen diese Attention-Mechanismen zur Extraktion von zusammenhängenden Merkmalen. Die Transformer Architektur wurde ursprünglich zur maschinellen Übersetzung eingesetzt und verwendet dabei die *stacked Self-Attention*, im Speziellen die *Multi-Head-Self-Attention*, und Fully-Connected-Schichten für den Encoder und Decoder [155]. Abbildung 2.19 zeigt die Architektur von Transformer-Netzen.

Die Sequenzanalyse bildet häufig die Grundlage für Workflowanalysen, die u. A. zur Untersuchung und Optimierung von Produktionsabläufen oder, wie im vorliegenden Fall, von Abläufen im OP-Umfeld eingesetzt werden.

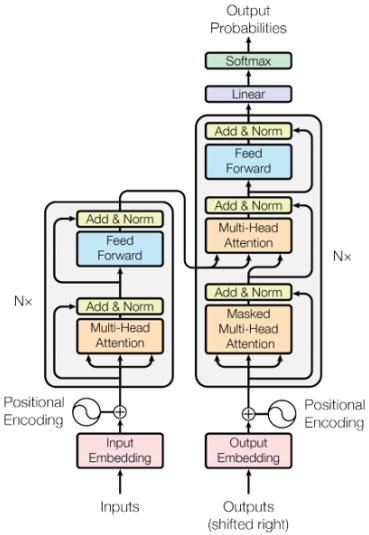


Abbildung 2.19: Architektur eines Transformer-Netzes (Quelle: [155])

2.2.4 Methoden und Metriken zur Evaluation von Systemen zur Kontexterfassung

Für die Evaluation von Kontexterfassungssystemen stehen eine Vielzahl an Methoden und Metriken zur Verfügung. Die in dieser Arbeit relevantesten werden in den folgenden Unterabschnitten für die weitere Verwendung erläutert.

Konfusionsmatrix: Die *Konfusionsmatrix* ist eine Möglichkeit zur grafischen Darstellung der Ergebnisse eines Klassifikators bezogen auf die Stichproben eines Datensatzes. Das Element kl_{ij} mit $i, j \in [1, N]$, wobei N der Anzahl an Klassen entspricht, zeigt dabei die Häufigkeit, in der eine Stichprobe aus der wahren Klasse i der Klasse j zugeordnet wurde [114]. Entsprechend zeigen die Werte auf der Diagonalen der Matrix die korrekt zugeordneten und die Werte außerhalb der Diagonalen die falschen Klassifizierungen [162]. Tabelle 2.1 stellt den erläuterten Zusammenhang beispielhaft für ein 2-Klassen-Problem dar.

		Wahrheit	
		Klasse 1	Klasse 2
Vorhersage	Klasse 1	kl_{11}	kl_{12}
	Klasse 2	kl_{21}	kl_{22}

Tabelle 2.1: Allgemeine Darstellung einer 2x2 Konfusionsmatrix

Dadurch lassen sich in Tabelle 2.1 die richtig klassifizierten Beispiele in *True Positive (TP)* und *True Negative (TN)* und die falsch klassifizierten Beispiele in *False Positive (FP)* und *False Negative (FN)* einordnen. TPs sind hierbei positiv klassifizierte Beispiele, die auch tatsächlich positiv sind und TNs negativ klassifizierte Beispiele, die auch tatsächlich negativ sind. Bei den Fehlerarten hingegen sind die FPs die positiv klassifizierten Beispiele, die tatsächlich negativ sind und FNs negativ klassifizierte Beispiele, die tatsächlich positiv sind [142]. Bezogen auf Klasse 1 sind folgende Zuordnungen in Tabelle 2.1 zutreffend:

- TP: kl_{11}
- TN: kl_{22}
- FP: kl_{12}
- FN: kl_{21}

Aus der Einordnung, die sich aus der Konfusionsmatrix ergibt, lassen sich unterschiedliche Metriken zur Evaluation von Klassifizierungsproblemen ableiten [32][122][98], von denen eine relevante Auswahl nachfolgend erläutert wird.

Genauigkeit: Die *Genauigkeit* (engl. *accuracy*) eines Modells ist ein einfaches und weit verbreitetes Maß für die Qualität der gesamten Klassifizierung und ist definiert als der Anteil der korrekt zugeordneten Fälle an der Grundgesamtheit einer Stichprobe [48]. Gleichung 2.1 stellt dar, wie die Genauigkeit berechnet wird:

$$\text{Genauigkeit} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Für ungleich verteilte Datensets wird die Genauigkeit allerdings stark durch häufiger vertretene Klassen beeinflusst, weshalb für deren Analyse meist zusätzliche Metriken eingesetzt werden.

Präzision: Die *Präzision* (engl. *precision*) ist der Anteil der vorhergesagten positiven Fälle, die sich als echte positive Fälle erweisen [122]. Gleichung 2.2 zeigt die zugehörige Berechnungsformel:

$$P = \frac{TP}{TP + FP} \tag{2.2}$$

Recall: Der *Recall*, auch Sensitivität oder True-Positive-Rate genannt, ist die Anzahl der positiven Vorhersagen im Verhältnis zu den wirklich positiven Fällen (vgl. Gleichung 2.3) [154].

$$R = \frac{TP}{TP + FN} \tag{2.3}$$

F1-Score: Der *F1-Score* ist der harmonische Mittelwert zwischen Präzision und Recall (vgl. Gleichung 2.4). Durch den Einbezug beider Gütemaße können deren spezifische Nachteile ausgeglichen werden, wodurch der F1-Score als einzelner Analysewert besser zum Vergleich verschiedener Klassifikatoren geeignet ist als die zuvor genannten Metriken [112]. Durch die Mittelung kann allerdings der Einfluss der einzelnen Vorhersagewerte nicht mehr genauer bestimmt werden.

$$\text{F1-Score} = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.4}$$

Funktionsgraphen: Die grafische Darstellung von Kurvenverläufen, die sich aus der Gegenüberstellung verschiedener, voneinander abhängiger Metriken ergeben, wird häufig genutzt, um einerseits die anwendungsspezifische optimale Konfiguration eines Modells zu finden und andererseits, um mehrere Modelle untereinander besser vergleichbar zu machen.

Eine häufig in anderen Arbeiten genutzte Variante, insb. bei ungleichmäßig verteilten Datensets, ist die *Precision-Recall-Kurve (PR-Kurve)*. Sie zeigt den Kompromiss zwischen Präzision und Recall für verschiedene Schwellwerte (im Gegensatz zu den Einzelwerten der bisher diskutierten Metriken). Eine große Fläche unterhalb der Kurve (engl. *Area under the Curve (AUC)*) steht sowohl für einen hohen Recall als auch für eine hohe Präzision, wobei, wie zuvor erläutert, eine hohe Präzision mit einer niedrigen Falsch-Positiv-Rate und ein hoher Recall mit einer niedrigen Falsch-Negativ-Rate verbunden ist. Hohe Werte für beide zeigen, dass der Klassifikator sowohl genaue Ergebnisse (hohe Präzision) als auch die Mehrheit aller tatsächlich positiven Ergebnisse (hoher Recall) liefert. Abbildung 2.20 zeigt beispielhaft eine solche PR-Kurve.

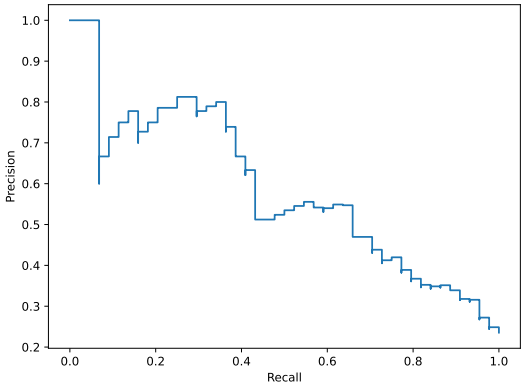


Abbildung 2.20: Beispiel einer PR-Kurve

Nach dem gleichen Prinzip können auch weitere Graphen zur Gegenüberstellung verschiedener Metriken aufgestellt werden. In dieser Arbeit werden zusätzlich

Precision-Kurven (P-Kurven), *Recall-Kurven (R-Kurven)* und *F1-Kurven* genutzt, die die jeweilige Metrik gegenüber der Erkennungskonfidenz darstellen. Auch diese Graphen dienen dem Finden geeigneter Schwellwerte, die ein möglichst robustes Modell versprechen.

Mean Average Precision: Die *Mean Average Precision (mAP)* ist eine der am weitesten verbreiteten Metriken zur Beurteilung von Modellen aus maschinellen Lernverfahren. In unterschiedlichen Anwendungsgebieten kann die mAP leicht unterschiedliche Bedeutungen haben. Hier soll eine möglichst allgemeingültige Definition dargestellt werden.

Wesentlicher Bestandteil zur Berechnung der mAP ist die *Average Precision (AP)*. Diese berechnet sich aus der AUC der PR-Kurve, was mathematisch im diskreten Fall wie folgt ausgedrückt werden kann:

$$AP = \sum_{s=1}^{S} P_s \cdot (R_s - R_{s-1}), \tag{2.5}$$

wobei P_s für die Präzision und R_s für den Recall bei Schwellwert s stehen und S die Anzahl der betrachteten Schwellwerte angibt.

Mit Hilfe von Gleichung 2.5 lässt sich die mAP aus dem arithmetischen Mittel der N APs aller untersuchten Klassen n berechnen:

$$mAP = \frac{1}{N} \sum_{n=0}^{N-1} AP_n \tag{2.6}$$

Rang-K Genauigkeit: Die *Rang-K Genauigkeit* (engl. *Rank-K Accuracy*, auch Top-K oder CMC-Top-K genannt) ist eine Metrik, welche häufig für die Evaluation von Re-Identifikationsalgorithmen verwendet wird [26], [146], [169], [174], [175]. Dafür werden die Ranglisten verwendet, die während des

Matchingprozesses (siehe Abschnitt 3.2.2) für verschiedene Queries erstellt wurden.

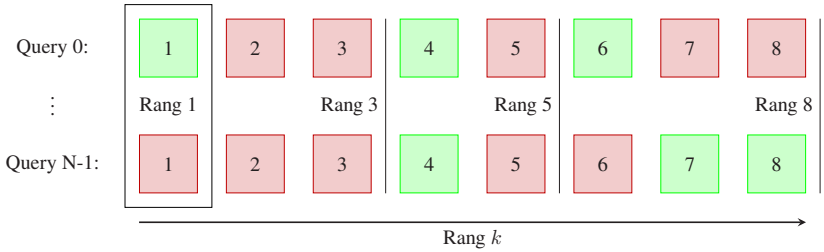


Abbildung 2.21: Berechnung der Rang-K Genauigkeit. Für die Queries 0 und N-1 sind die Ranglisten des Matchings dargestellt. Die Ranglisten laufen von links (Rang k=1) nach rechts (Rang k=8). Die farbigen Kästen einer Rangliste repräsentieren die zugeordneten Aufnahmen aus der Galerie. Bei einer korrekten Zuordnung ist der Kasten grün eingefärbt, bei nicht korrekter Zuordnung rot.

Dabei wird überprüft, für wie viele der N betrachteten Queries eine korrekte Zuordnung innerhalb eines bestimmten Rangs k vorliegt. Abbildung 2.21 verdeutlicht die Funktionsweise der Metrik. Mathematisch lässt sie sich, wie in Gleichung 2.7 dargestellt, beschreiben.

$$R_k = \frac{1}{N} \sum_{q=0}^{N-1} m_q(k) \tag{2.7}$$

Dabei ist N die Anzahl der betrachteten Queries q. Des Weiteren ist $m_q(k)$ eine Indikatorfunktion, die besagt, ob für das jeweilige Query q eine korrekte Zuordnung innerhalb des k-ten Rangs vorliegt oder nicht. Liegt eine korrekte Zuordnung in den ersten k Rängen vor, dann ist $m_q(k) = 1$, ansonsten gilt $m_q(k) = 0$. Für das in Abbildung 2.21 gezeigte Beispiel ergeben sich, für $N = 2$ mit den zwei dargestellten Queries, die Rang-1, -3, -5 und -8 Genauigkeiten wie folgt:

$$R_1 = \frac{1}{2} \cdot (1 + 0) = 50\%$$

$$R_3 = \frac{1}{2} \cdot (1 + 0) = 50\%$$

$$R_5 = \frac{1}{2} \cdot (1 + 1) = 100\%$$

$$R_8 = \frac{1}{2} \cdot (1 + 1) = 100\%$$

Phasen-Zeit-Diagramm: Bei der Segmentierung einzelner Phasen einer OP kommen in dieser Arbeit zur Visualisierung der Ergebnisse sog. *Phasen-Zeit-Diagramme* zum Einsatz. Dabei wird für jeden Frame des untersuchten Videos zum einen die Ground-Truth-Phase und zum anderen die prädizierte Phase farblich dargestellt. Somit kann die Qualität der Erkennung über den gesamten zeitlichen Verlauf erfasst werden. Abbildung 2.22 skizziert beispielhaft ein Phasen-Zeit-Diagramm zur Veranschaulichung.

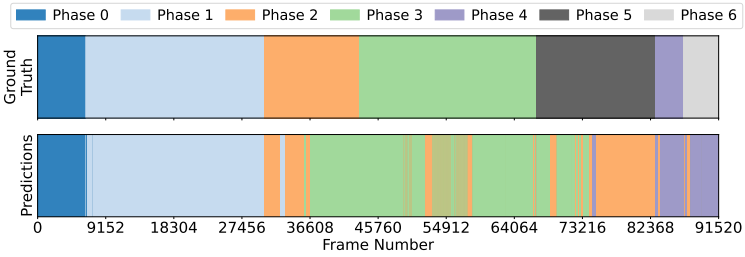


Abbildung 2.22: Beispiel eines Phasen-Zeit-Diagramms

3 Stand der Wissenschaft und Technik

In diesem Kapitel wird der für die vorliegende Arbeit relevante Stand der Wissenschaft und Technik dargestellt. Hierfür werden zunächst Deep-Learning-Ansätze zur Objekterkennung diskutiert. Im Anschluss werden Methoden zur Erkennung und zum Tracking menschlicher Posen als Grundlage der Verhaltensanalyse betrachtet. Danach erfolgt eine Erörterung bzgl. der Erkennung spezifischer OP-Phasen und der daraus ableitbaren Erkenntnisse zum Workflow innerhalb des Operationssaals, bevor abschließend verschiedene Datensätze für die datengetriebene Entwicklung in der vorliegenden Anwendungsdomäne erläutert werden.

3.1 Objekterkennung

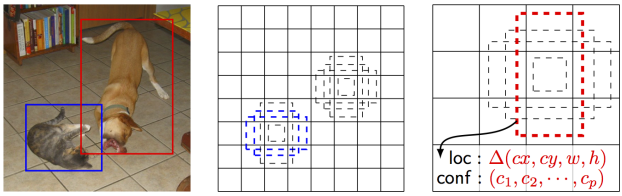
Wie bereits im Grundlagenkapitel (siehe Kapitel 2) erläutert, werden im Verlauf dieser Arbeit lediglich Deep Learning-basierte Ansätze zur Objekterkennung betrachtet. Aus diesem Grund beschränkt sich die Betrachtung zum Stand der Technik ebenfalls auf solche Methoden. Im Folgenden wird eine Auswahl an Objekterkennungsmodellen mit Relevanz für die vorliegende Arbeit betrachtet.

Faster R-CNN ist eine Erweiterung des *Region-based Convolutional Neural Network (R-CNN)* Ansatzes zur Bounding-Box-Objekterkennung. Dieser nutzt vorgeschlagene Objektregionen und Faltungsnetzwerke, die unabhängig voneinander auf verschiedenen *Regions of Interest (ROIs)* ausgewertet werden [58]. Faster R-CNN lernt dabei einen *Attention*-Mechanismus mit einem sog. *Region*

Proposal Network (RPN) [130]. Das Gesamtmodell besteht aus zwei Stufen. In der ersten Phase, dem RPN, werden Kandidaten für Bounding Boxes vorgeschlagen. In der zweiten Phase, die im Wesentlichen ein Fast R-CNN [57] ist, werden mithilfe von *RoIPool* Merkmale aus jeder vorgeschlagenen Bounding Box extrahiert und eine Klassifizierung und Bounding-Box-Regression durchgeführt. Die von beiden Stufen verwendeten Merkmale können zur schnelleren Inferenz gemeinsam genutzt werden [66]. Die Trainings- und Testbilder müssen jeweils nur eine einzige Skalierung aufweisen, was positiv auf die Ausführungsgeschwindigkeit einwirkt. Durch die Nutzung von *Multi-scale Anchors* in den RPNs kann dennoch eine Skalierungsinvarianz erreicht werden. Somit sind insgesamt gute Erkennungsraten möglich [131]. Die Multi-Stage-Architektur und die Nutzung von Ankern bringen allerdings Overhead mit sich, der dazu führt, dass Faster R-CNN erhebliche Nachteile bei der Inferenzgeschwindigkeit im Vergleich zu anderen Ansätzen hat.

Der *Single-Shot-Multibox-Detektor (SSD)* ist eine Methode zur Erkennung von Objekten in Bildern mithilfe eines einzigen tiefen neuronalen Netzes. Dieser Ansatz diskretisiert den Ausgaberaum von Bounding Boxes in eine Reihe von Standardboxen mit unterschiedlichen Seitenverhältnissen und Skalen pro Feature Map Position. SSD ist im Vergleich zu Methoden, die Objektvorschläge erfordern, einfach, da es die Generierung von Vorschlägen und die anschließende Neuabtastung von Pixeln oder Merkmalen vollständig eliminiert und alle Berechnungen in einem gemeinsamen Netzwerk bündelt. Dadurch ist SSD einfach zu trainieren und kann problemlos in Systeme integriert werden, die eine Erkennungskomponente benötigen. Der SSD-Ansatz basiert auf einem *Feed-Forward* CNN, das eine in der Größe festgelegte Sammlung von Bounding Boxes und Konfidenzwerten für das Vorhandensein von Objektklasseninstanzen in diesen Boxen erzeugt. Danach werden mittels *Non-Maximum Suppression (NMS)* die endgültigen Erkennungen extrahiert [107].

In Abbildung 3.1 ist die Funktionsweise eines SSD-Netzwerks skizziert. Abbildung 3.1a zeigt dabei ein Eingabebild mit den zugehörigen Ground-Truth-Boxen für jedes zu erkennende Objekt für das Modelltraining. Nach dem Faltungsprinzip wird ein kleiner Satz (z. B. 4) von Standardboxen mit unterschiedlichen



(a) Bild mit Ground-Truth-Boxen **(b)** 8×8 Feature Map **(c)** 4×4 Feature Map

Abbildung 3.1: Funktionsweise des SSD-Frameworks nach [107].

Seitenverhältnissen an jeder Stelle in mehreren Feature Maps mit unterschiedlichen Skalierungen evaluiert (z.B. 8×8 in Abbildung 3.1b und 4×4 in Abbildung 3.1c). Für jede Standardbox werden sowohl die Verschiebungen der Bounding Boxes als auch die Konfidenzwerte für alle Objektkategorien $(c_1, c_2, \ldots, c_p)$ vorhergesagt. Zum Zeitpunkt des Trainings werden diese Standardboxen zunächst mit den Ground Truth Boxen verglichen. Im Beispiel sind zwei Standardboxen der Katze (blau markiert in Abbildung 3.1b) und eine dem Hund zugeordnet (rot markiert in Abbildung 3.1c). Diese werden als positiv, die übrigen als negativ behandelt [107].

Bei *You Only Look Once (YOLO)* handelt es sich um einen Ansatz zur Objekterkennung, der als Regressionsproblem auf räumlich getrennte Bounding Boxes und damit verbundene Klassenwahrscheinlichkeiten betrachtet wird. Wesentliches Merkmal der YOLO-Modellwelt ist die hohe Detektionsgeschwindigkeit, die Erkennungen in Echtzeit erlaubt. Im Gegensatz zu den meisten Ansätzen, die früher entwickelt wurden, führt YOLO die Merkmalsextraktion, Bounding-Box-Vorhersage, NMS und die darauf aufbauende Prädiktion auf dem gesamten Bild parallel innerhalb eines einzigen CNN durch [128].

Die von Redmon et al. ursprünglich entwickelte Funktionsweise (auch bekannt als *YOLOv1*) ist in Abbildung 3.2 skizziert. Folgende Arbeitsschritte sind dafür notwendig [128]:

1. Das Eingangsbild wird in $S \times S$ gleich große Zellen eingeteilt.

2. In jeder Zelle werden nun B Bounding Boxes und der jeweilige *Box Confidence Score* berechnet. Die Zelle, die die Mitte einer Bounding Box enthält, bestimmt deren Klasse.

3. Für alle Zellen wird die *bedingte Klassenwahrscheinlichkeit* für jede Klasse berechnet.

4. Durch Multiplikation der bedingten Klassenwahrscheinlichkeiten und der individuellen Box-Konfidenz-Vorhersagen entsteht schließlich ein *klassenspezifischer Konfidenzwert* für jede Bounding Box.

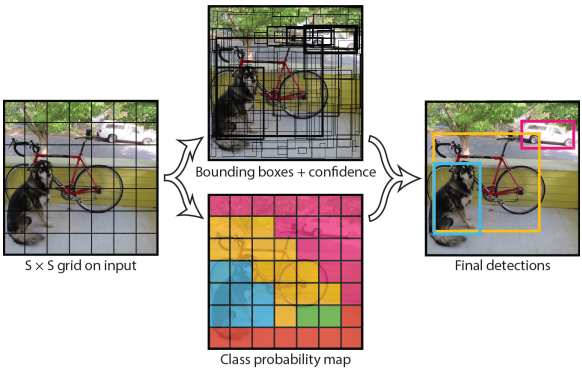


Abbildung 3.2: Schematische Darstellung des Ablaufs der YOLO Objekterkennung [128].

Die Vorteile, die YOLO in seiner ursprünglichen Form mitbringt, sind dessen Geschwindigkeit, geringer Background-Error und eine gute Generalisierbarkeit. Dagegen ist die Erkennungsperformance schlechter als der damalige Stand der Technik, was insb. an der Anfälligkeit für Lokalisierungsfehler und der schlechten Erkennbarkeit von kleinen Objekten, die nahe beieinander liegen, lag.

YOLO wurde seit seiner ursprünglichen Veröffentlichung mehrfach weiterentwickelt, um unterschiedliche konzeptionelle Schwachstellen zu verbessern. Dadurch konnte sowohl die Erkennungsqualität als auch die Geschwindigkeit

weiter gesteigert werden. Bemerkenswert dabei ist, dass nicht alle Modellvarianten, die unter dem Namen *YOLO* veröffentlicht wurden und auch in der Community verbreitet sind, von den ursprünglichen Entwicklern stammen. Dies führte dazu, dass teilweise neue Versionen in sehr kurzen zeitlichen Abständen entstanden und diese unterschiedliche Backends einsetzen. Die folgende Zusammenfassung der zum Zeitpunkt der Entstehung dieser Arbeit relevantesten Modellvarianten ist weitestgehend den Erläuterungen von Jocher and Chaurasia in [84] entnommen:

- **YOLOv2** (2016): Verbessert das ursprüngliche Modell durch die Einbeziehung von *Batch Normalization*, *Anker Boxen* und *Dimension Clusters* [127]
- **YOLOv3** (2018): Steigerung der Modell-Performance durch die Nutzung eines effizienteren Backbone-Netzwerks, mehr Ankern und dem Einsatz von *Spatial Pyramid Pooling (SPP)* [126]
- **YOLOv4** (2020): Einführung u. A. von *Mosaik Data Augmentation*, einem neuen ankerlosen Detection Head und einer neuen Verlustfunktion [13]
- **YOLOv5** (2020): Weitere Verbesserung der Leistung des Modells und neue Funktionen wie Hyperparameter-Optimierung, integriertes Experimententracking und automatischer Export in gängige Exportformate [152]
- **YOLOv6** (2022): Nutzung eines neuen Backbones mit höherem Parallelisierungsgrad und entkoppeltem Detection Head für Klassenkonfidenz und Lokalisierung zusammen mit Intersection over Union (IoU), neuer Verlustfunktion sowie neuer Quantisierungsstrategien [102]
- **YOLOv7** (2022): Ergänzt zusätzliche Tasks wie Posenerkennung und *Instance Segmentation*, Architekturänderungen und Nutzung zusätzlicher *Bag-of-Freebies* zur Performancesteigerung ohne die Inferenzzeit zu beeinflussen [160]

- **YOLOv8** (2023): Weiterentwickelte Backbone- und Neck-Architektur, Erweiterung um zusätzliche Computer-Vision-Aufgaben, entkoppelter Detection Head für unabhängige Berechnung der Klassenkonfidenz, Lokalisierung und Klassifikation [85]
- **YOLO-NAS** (2023): Automatisches Architekturdesign durch *AutoNAC*, *Quantization Aware Modules* zur Verminderung von Genauigkeitsverlusten durch Quantisierung und *hybride Quantisierungsmethoden* für bessere Balance zwischen Latenz und Genauigkeit führen insgesamt zur Verbesserung der Detektion kleiner Objekte und der Lokalisierung [3].

Terven and Cordova-Esparza geben in [148] einen detaillierteren Überblick über die verschiedenen YOLO-Iterationen und deren spezifische Eigenschaften.

Viele der unterschiedlichen YOLO-Versionen werden in verschiedenen Varianten angeboten, die üblicherweise durch verschiedene Modellgrößen eine bessere Auswahl auf die verfügbaren Rahmenbedingungen ermöglichen. Dies geht im Regelfall mit entsprechender Abwägung zwischen Geschwindigkeit und Erkennungsqualität einher. In vereinzelten Fällen werden auch Varianten mit unterschiedlichen Architekturvariationen angeboten. So gibt es für YOLOv3 nicht nur eine kleinere *tiny*-Variante, sondern darüber hinaus auch Varianten mit oder ohne SPP.

Ein beliebtes Framework, das verschiedene YOLO-Versionen vortrainiert integriert und auch darauf basierendes Finetuning ermöglicht, wird von dem Unternehmen *Ultralytics* bereitgestellt [153].

Bei *Real-Time Detection Transformer (RT-DETR)* handelt es sich um einen transformerbasierten Ansatz einer echtzeitfähigen End-to-End Objekterkennung [110]. Es nutzt die Leistungsfähigkeit von Vision Transformers (ViTs) zur effizienten Verarbeitung von multi-scale Features durch Entkopplung von Interaktion innerhalb der gleichen Skalierungen und Fusion unterschiedlicher Skalierungen. RT-DETR ist hochgradig anpassungsfähig und unterstützt eine flexible Anpassung der Inferenzgeschwindigkeit unter Verwendung verschiedener Decoderschichten ohne den Bedarf eines neuen Trainings [84].

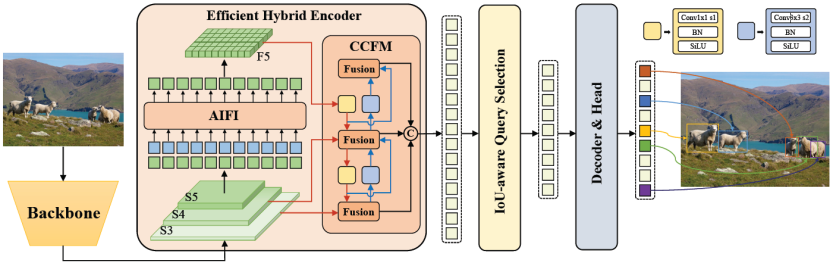


Abbildung 3.3: Schematische Darstellung der RT-DETR-Architektur (Quelle: [110])

Abbildung 3.3 skizziert den Aufbau eines RT-DETR-Netzwerks. Es zeigt die letzten drei Stufen des Backbones (S3, S4, S5) als Eingabe des Encoders. Der effiziente Hybrid-Encoder wandelt Multiskalenmerkmale in eine Sequenz von Bildmerkmalen um. Dies geschieht durch die Interaktion von Intraskalenmerkmalen (*intrascale feature interaction (AIFI)*) und das skalenübergreifende Merkmals-Fusionsmodul (*cross-scale feature-fusion module (CCFM)*). Die Funktionsweise des Hybrid-Encoders verhindert redundante Berechnungen und ermöglicht somit eine effiziente Verarbeitung von Merkmalen in verschiedenen Skalierungen. Die *IoU-aware query selection* wird eingesetzt, um eine feste Anzahl von Bildmerkmalen auszuwählen, die als initiale Objektabfrage für den Decoder dienen. Schließlich optimiert der Decoder mit zusätzlichen *prediction heads* iterativ die Objektabfragen, um Bounding Boxes und Konfidenzwerte zu erzeugen [110]. Für die Objekterkennung wird kein Postprocessing benötigt, was dafür sorgt, dass die Inferenzzeit nicht verzögert und somit die Echtzeitfähigkeit ermöglicht wird.

3.2 Erkennung menschlicher Bewegung

3.2.1 Posenerkennung & -tracking

Für die Posenerkennung stehen verschiedene Bibliotheken zur Verfügung, darunter auch einige Open-Source-Projekte. Einige Beispiele hierfür sind: *OpenPose*

[24, 23, 139, 164], *AlphaPose* [45, 44, 103], *DeepCut* [76, 121] oder *RMPE* [44]. Aufgrund einer guten Dokumentation und der weiten Verbreitung der Bibliothek wurde im Folgenden zunächst OpenPose verwendet.

OpenPose ist eine Bibliothek zur 2D Multi-Person Pose Estimation. Es erhält als Input einen Frame eines Videos, aus welchem mithilfe eines CNNs sogenannte *Confidence Maps* erzeugt werden. Confidence Maps zeigen die Aufenthaltswahrscheinlichkeit von Keypoints im betrachteten Bild (vgl. Abbildung 3.4 (b)). Zusätzlich werden Vektorfelder erzeugt, welche die Lage und Orientierung von Körperteilen beschreiben. Diese *Part Affinity Fields* (vgl. Abbildung 3.4 (c)) ermöglichen eine bessere Zuordnung von Keypoints zu einzelnen Personen. Die Zuordnung der Keypoints kann als ein Problem der Graphentheorie (*bipartites Matchingproblem*) beschrieben werden (vgl. Abbildung 3.4 (d)). Hierfür gibt es entsprechende Lösungsverfahren, die hier nicht weiter erläutert werden. Das Ergebnis ergibt von jeder Person ein Skelett, welches die Körperhaltung zweidimensional nachbildet (vgl. Abbildung 3.4 (e)) [24].

Abbildung 3.4: Funktionsweise von OpenPose nach [24]

Da im Laufe der Arbeit zur Verhaltensanalyse aus Deckenkameras Probleme beim Tracking mit OpenPose festgestellt wurden (vgl. Kapitel 5), wurde stattdessen im späteren Verlauf AlphaPose eingesetzt. AlphaPose ist ein System zum akkuraten Erkennen und Tracken von Posen in Echtzeit. Dafür werden in [45] verschiedene neuartige Techniken erläutert: Die *Symmetrische integrale Keypoint-Regression (SIKR)* erlaubt eine schnelle und feine Lokalisierung, *parametrische Pose Non-Maximum-Suppression (P-NMS)* zur Eliminierung redundanter Erkennungen von Menschen und *Pose Aware Identity Embedding*

zur Kombination von Posenschätzung und -tracking. Während des Trainings kommen ein *Part-Guided Proposal Generator (PGPG)* und eine *Multi-Domain Knowledge Distillation* zum Einsatz, um die Genauigkeit weiter zu verbessern.

3.2.2 Re-Identifikation von Personen

Die meisten bisherigen Arbeiten zur Re-ID konzentrieren sich auf die Verbesserung der Re-Identifizierung von Fußgängern [129], daher wurden in der Regel Personen in Alltagskleidung beobachtet. Die Verwendung von Alltagskleidung führt häufig zu einer eindeutigen Identifizierbarkeit von Personen, bspw. durch Merkmale wie der Statur, Frisur oder Kleidungsstil. Dies vereinfacht die Aufgabe der Re-Identifizierung erheblich. Darüber hinaus findet die Re-ID von Personen auch an Orten und Domänen Anwendung, an denen Menschen normalerweise keine Alltagskleidung tragen, z. B. im Sport, beim Militär oder im medizinischen Umfeld. Stattdessen tragen sie meist einheitliche Trikots oder Arbeitskleidung. Da die einzelnen Kleidungsstücke dadurch oft sehr ähnlich aussehen, geht das individuelle Aussehen der Personen verloren und sie ähneln sich sehr stark. Dies kann zu Problemen bei der Re-Identifikation führen. Der Stand der Technik hierzu fokussiert sich fast ausschließlich auf Deep Learning-basierte Ansätze, weshalb deren grundlegende Funktionsweise hier näher beschrieben wird.

In der Regel bestehen Re-Identifikationsysteme, welche mit Deep-Learning arbeiten, aus einer Merkmalsextraktion, einer speziellen Verlustfunktion und dem Matchingprozess [166]. Abbildung 3.5 stellt diesen Zusammenhang und die Unterscheidung zwischen dem Training eines solchen Modells und dessen Validierung dar. Die Funktionsweise der einzelnen Komponenten wird im Folgenden genauer beschrieben.

Das System erhält zum Training des Re-ID-Modells zunächst annotierte Trainingsdaten. Die Annotationen sind in diesem Fall eindeutige IDs, die den sichtbaren Personen zugeordnet sind. Für die Validierung erhält das System mehrere Queries, welchen eine ID zugeordnet werden soll, sowie eine Galerie

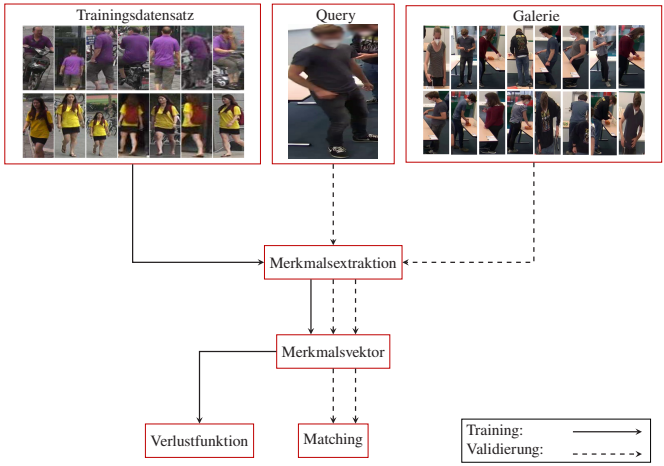


Abbildung 3.5: Ablauf von Training und Validierung für Re-Identifikationsmodelle. Ausgehend von den Datensätzen werden für Training und Validierung eines Re-Identifikationmodells unterschiedliche Pfade durchlaufen (Bildquelle Trainingsdatensatz: [171]).

mit Einzelbildern, die zum Matching der Queries verwendet werden. Die Personen in den Trainingsdaten müssen nicht dieselben Personen sein, die auch als Query oder in der Galerie vorkommen. Eine Zuordnung von Personen ist auch dann möglich, wenn diese nicht im Trainingsdatensatz enthalten sind. Damit eine korrekte Zuordnung der ID möglich ist, muss allerdings mindestens eine zweite Aufnahme des Queries in der Galerie vorhanden sein. Diese Aufnahmen müssen nicht identisch sein, sondern können sich hinsichtlich Perspektive, Umgebung oder ausgeübter Tätigkeit unterscheiden. Für das Query sowie für die Bilder der Galerie können Annotationen vorliegen, diese sind allerdings nur für die Validierung des Re-Identifikationsalgorithmus relevant. Für die Zuordnung eines Queries zur Galerie werden keine Annotationen benötigt.

Der erste Schritt der Re-ID von Personen ist die Merkmalsextraktion. Ziel dabei ist es, ähnlich wie bei Klassifikationsmodellen, aus den Einzelbildern der Personen unterscheidbare Merkmale zu extrahieren, welche die Personen möglichst eindeutig beschreiben [166], [161]. Viele Re-Identifikationsmodelle verwenden deshalb als Grundbaustein Architekturen, die ursprünglich für die

Klassifikation von Bildern entwickelt wurden [166]. Häufig handelt es sich dabei um Variationen eines CNN. Zhou et al. verwenden bspw. ein CNN mit verschiedenen Faltungskernen, um unterschiedlich große Merkmale zu erfassen [174]. Sun et al. teilen das Bild in mehrere horizontale Schichten auf, um verschiedene Körperabschnitte einzeln zu betrachten. Anschließend verwenden sie ein CNN, um aus den verschiedenen Schichten Merkmale zu extrahieren [146]. Weitere Beispiele finden sich in [26], [169] oder [175].

Die *Verlustfunktion* ist ein elementarer Bestandteil zur Quantifizierung der Güte des Trainings von Deep-Learning-Modellen. Beim Training von Re-ID-Modellen werden spezielle Verlustfunktionen eingesetzt, um die Erstellung des Merkmalsvektors zu optimieren. Im Folgenden werden dafür der *Identity Loss* und der *Triplet Loss* vorgestellt.

Der *Identity Loss* nutzt Konzepte der (Bild-)Klassifikation für das Training von Re-ID-Modellen. Dabei stellt jede ID einer Person eine eigene Klasse dar [166]. Der Identity Loss wird auch als *Classification Loss* oder als *Softmax Loss* bezeichnet.

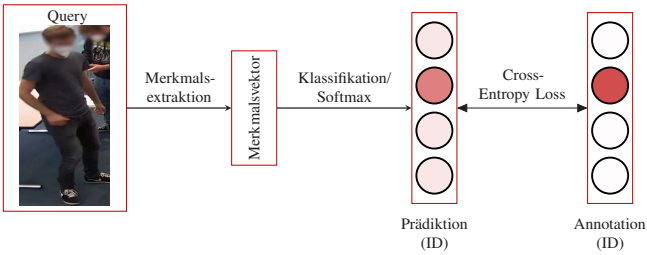


Abbildung 3.6: Training eines Re-Identifikationsmodells mittels Identity Loss. Dieser setzt sich zusammen aus einer Klassifikation und anschließender Verwendung des Cross-Entropy Loss zwischen der zugeordneten ID (Prädiktion) und der tatsächlich annotierten ID.

Abbildung 3.6 zeigt die Funktionsweise des Identity Loss. Mithilfe einer Softmax-Funktion wird eine Wahrscheinlichkeitsverteilung über die Zuordnung des Merkmalsvektors zu einer bestimmten ID erstellt. Für ein Bild x_i und die zugehörige ID y_i ist $p(y_i|x_i)$ die Wahrscheinlichkeit, dass die Person

im Bild, der Klasse y_i zugeordnet wurde. Als eigentliche Verlustfunktion wird anschließend der *Cross-Entropy Loss* verwendet, um die Abweichung zwischen der Wahrscheinlichkeitsverteilung und der tatsächlichen ID zu bestimmen. Der *Cross-Entropy Loss* ergibt sich wie folgt [166]:

$$L_{id} = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i) \tag{3.1}$$

wobei n die Anzahl der Trainingssamples in einem Batch beschreibt.

Die Verwendung von *Triplet Loss* als Verlustfunktion zielt darauf ab, dass Bilder derselben Person einen ähnlichen Merkmalsvektor hervorrufen und sich gleichzeitig Merkmalsvektoren von verschiedenen Personen stärker unterscheiden. Zur Berechnung des Triplet Loss werden drei Bilder aus den Trainingsdaten benötigt: ein Referenzbild (*Anker*), ein Bild mit derselben Person wie im Referenzbild (*Positiv*) und ein Bild mit einer beliebigen anderen Person (*Negativ*). Während des Trainings werden also jeweils drei Aufnahmen betrachtet, wobei die Verlustfunktion hinsichtlich der ID des Ankers bzw. des Positivs optimiert wird [166]. Zur Veranschaulichung ist ein Beispiel in Abbildung 3.7 dargestellt.

Gleichung 3.2 beschreibt den Triplet Loss für die Merkmalsvektoren x_i (Anker), x_j (Positiv) und x_k (Negativ). Dabei ist ρ ein gewünschter Mindestabstand zwischen den Klassen und $d(\cdot)$ der euklidische Abstand zwischen zwei Merkmalsvektoren [166].

$$L_{triplet}(i,j,k) = \max(\rho + d_{ij} - d_{ik}, 0) \tag{3.2}$$

Das Ziel während des Trainingsprozesses ist die Minimierung der Verlustfunktion $L_{triplet}(i,j,k)$. Diese wird minimal, falls $\rho + d_{ij} < d_{ik}$ gilt. Das bedeutet, dass sich die Verlustfunktion minimiert, wenn sich der Abstand d_{ij} zwischen zwei Merkmalsvektoren derselben Person verkleinert und der Abstand d_{ik} zwischen Merkmalsvektoren verschiedener Personen vergrößert.

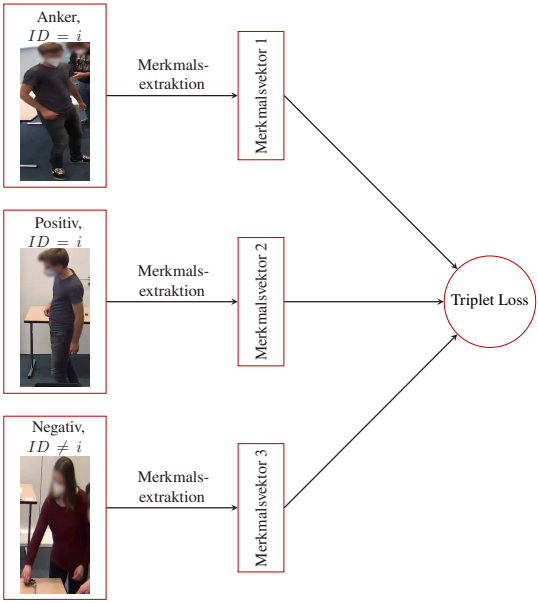


Abbildung 3.7: Für die Berechnung des Triplet Loss werden drei Aufnahmen (Anker, Positiv, Negativ) verwendet. Für jede dieser Aufnahmen wird ein Merkmalsvektor erstellt, welcher in die Berechnung mit einfließt.

Das Matching ist der Prozess, bei welchem letztendlich einem Query eine ID zugeordnet wird. Während des Matchingprozesses wird für alle Merkmalsvektoren aus der Galerie die Ähnlichkeit zum Merkmalsvektor des Queries bestimmt. Basierend darauf wird eine Rangliste der Merkmalsvektoren erstellt, welche nach absteigender Ähnlichkeit geordnet ist. Die erzeugte Rangliste wird verwendet, um dem Query eine ID zuzuordnen [9]. Zur Bestimmung der Ähnlichkeit zweier Merkmalsvektoren können beliebige Distanzmetriken, wie bspw. die euklidische Distanz, die Bhattacharyya Distanz oder die Kosinusdistanz, verwendet werden. Dabei sind zwei Merkmalsvektoren umso ähnlicher, je geringer die Distanz zwischen ihnen ist.

Unter Verwendung eines der Ähnlichkeitsmaße wird eine Rangliste der Merkmalsvektoren bzw. der zugehörigen Galeriebilder erstellt. Das Bild auf dem ersten Rang ist jenes, welches den Merkmalsvektor mit der geringsten Distanz

zum Merkmalsvektor des Queries besitzt. Abschließend wird die ID des ersten Rangs aus den Annotationen entnommen und dem Query zugeordnet. Im Falle einer korrekten Zuordnung handelt es sich bei dem Query und dem Bild auf dem ersten Rang um dieselbe Person. Andernfalls ist die Zuordnung falsch. Um das Matching zu verbessern, können verschiedene Methoden des *Re-Rankings* verwendet werden. Dabei wird die erstellte Rangliste abhängig von Ähnlichkeiten zwischen verschiedenen Galeriebildern neu angeordnet. Ye et al. liefern einen Überblick über verschiedene Methoden zur Optimierung der Rangliste [166].

Im folgenden Abschnitt werden beispielhaft einige Modelle au dem Stand der Technik dargelegt.

Fu et al. verwenden in [50] unüberwachtes Training auf einem großen, nicht annotierten Datensatz zur Re-Identifikation von Personen. Sie erreichen eine mAP von 92 Prozent und eine Rang-1-Genauigkeit von 97 Prozent für den Market1501-Datensatz. Wieczorek et al. können diese Ergebnisse übertreffen, indem sie eine neue Verlustfunktion namens Centroid Triplet Loss verwenden. Sie erreichen 98,3 Prozent mAP und 98 Prozent Rang-1-Genauigkeit auf dem Market1501-Datensatz sowie 96,1 Prozent mAP und 95,6 Prozent Rang-1-Genauigkeit auf dem DukeMTMC-Datensatz [165]. Diese Ergebnisse wurden mit Bildern von Menschen in Alltagskleidung erzielt. Im Gegensatz dazu wird in der vorliegenden Arbeit medizinische Kleidung anstelle von Alltagskleidung verwendet. Obwohl die Auswirkungen speziell von medizinischer Kleidung in der Literatur noch nicht tiefergehend untersucht wurden, gibt es Arbeiten, die sich mit ähnlichen Szenarien befassen.

Bialkowski et al. untersuchen bspw. die Re-ID von Personen auf Videodaten von Mannschaftssportveranstaltungen. Da bei Mannschaftssportarten häufig einheitliche Sportkleidung getragen wird, ist das Erscheinungsbild der verschiedenen Personen hier ebenfalls sehr ähnlich. Aufgrund der unzureichenden Auflösung der in dieser Arbeit genutzten Bilder können Gesichtsmerkmale und Trikotnummern nicht zur eindeutigen Identifizierung der Personen verwendet werden. Daher verwenden die Autoren gruppendynamische Informationen zur Optimierung der Re-ID. Die Grundidee der Arbeit besteht darin, dass Menschen soziale

Wesen sind und sich im Allgemeinen als kollektive Gruppe bewegen. Sie lernen eine lineare Abbildungsfunktion, die jedem Spieler eine Rolle und Position innerhalb der Gruppenstruktur zuweist. Die Ergebnisse zeigen, dass Gruppeninformationen die Re-ID von Personen in einer Sportumgebung im Vergleich zu Methoden, die sich nur auf Erscheinungsmerkmale stützen, verbessern können.

Yin et al. beschreiben ein Problem in der Tatsache, dass die meisten Wiedererkennungsmethoden davon ausgehen, dass das visuelle Erscheinungsbild einer Person und insbesondere die Farbe ein relevantes Unterscheidungsmerkmal ist. Sie bezeichnen dieses Problem als *fine-grained person re-identification (FGPR)*. Um FGPR zu untersuchen, erstellen die Autoren einen Datensatz, indem sie Videos von Personen mit einheitlicher Kleidung sammeln. Anhand der Farbe ihrer Hemden werden die Personen in drei Gruppen eingeteilt: die blaue, weiße und grüne Gruppe. Die Autoren gehen davon aus, dass jede Person über eigene einzigartige Posen und Bewegungsmerkmale verfügt. Sie kombinieren daher globale Erscheinungsmerkmale und Bewegungsmerkmale, um *bewegungsabhängige lokale dynamische Posenmerkmale* und *gelenkspezifische lokale dynamische Posenmerkmale* zu generieren. Mit dieser Methode erreichen sie 88,4 Prozent mAP und 87,1 Prozent Rang-1-Genauigkeit auf ihrem eigenen Datensatz.

Im späteren Verlauf der vorliegenden Arbeit wird die speziell für *Deep Learning person re-identification* entwickelten Bibliothek *Torchreid* genutzt [173]. Eine Auswahl der dort implementierten Modelle wird im Folgenden erläutert.

OSNet verwendet *omni-scale features*, die eine Mischung von Merkmalen aus verschiedenen Skalen enthalten. Damit werden sowohl Merkmale des gesamten Körpers als auch kleine Details erfasst. Sie erreichen damit auf Market1501 eine mAP von 84,9 und eine Rank-1-Genauigkeit von 94,8 [174]. *OSNet-AIN* basiert auf OSNet, verwendet aber zusätzlich eine *instance normalization*. Diese minimiert datensatzspezifische Merkmale, die durch individuelle Umgebungen oder Lichtverhältnisse verursacht werden. Dadurch wird insb. die domänenübergreifende Performance verbessert [175]. *PCB* versucht körperteilspezifische Merkmale zu erlernen, indem die Bilder in mehrere Teile zerlegt werden. *Refined part pooling* ermöglicht die Zuordnung von Ausreißern zu den richtigen

Körperteilen zuzuordnen. Damit kann eine mAP von bis zu 81,6 und eine Rang-1-Genauigkeit von 93,8 auf Market1501 erreicht werden [146]. *ResNet-Mid* [169] und *MLFN* [26] verwenden beide Merkmale aus verschiedenen semantischen Ebenen und argumentieren, dass Merkmale aus verschiedenen Stufen des neuronalen Netzes Informationen mit spezifischen Abstraktionsebenen enthalten. ResNet-Mid resultiert damit in einer maximalen mAP von 82,37 und Rang-1-Genauigkeit von 93,32 auf Market1501 [169], MLFN kommt auf ähnliche Ergebnisse mit einer mAP von 82,4 und einer Rang-1-Genauigkeit von 92,3 [26].

3.3 Workflowanalyse im OP

Die Verwendung von Videoszenen zur automatischen Erkennung chirurgischer Phasen stellt eine große Herausforderung dar. Zum einen herrscht oftmals eine nur geringe Inter-Klassen-Varianz zwischen den verschiedenen Phasen bei gleichzeitig großer Intra-Klassen-Varianz. Zum anderen kommt es aufgrund der Kamerabewegung und der während des Eingriffs entstehenden Artefakte wie bspw. Rauch oder Blutungen zu einer starken Unschärfe der Szene, was die Erkennung erschwert. Darüber hinaus kann es bei komplexen chirurgischen Eingriffen vorkommen, dass das Kamerabild nicht immer den relevanten Ausschnitt der chirurgischen Szenen zeigt, wodurch zusätzliches Rauschen und Artefakte in die aufgezeichneten Videos gelangen.

Im vorliegenden Abschnitt wird der Stand der Technik in Bezug auf die automatisierte Phasenerkennung von OPs dargelegt. Der Schwerpunkt liegt dabei, soweit es für die Methode relevant ist, auf Arbeiten, die sich auf laparoskopische chirurgische Eingriffe fokussieren.

Frühe Arbeiten zur Erkennung von Operationsphasen aus chirurgischen Videos basieren hauptsächlich auf manuell erstellten Merkmalen, bspw. Pixelwerten, Intensitätsgradienten, räumlich-zeitlichen Merkmalen und Merkmalen wie Farbe, Textur und Form. Parallel dazu gibt es auch Arbeiten, die lineare statistische

Modelle für die Erfassung der zeitlichen Informationen von OP-Videos verwenden. Beispiele dafür sind das *Left-Right Hidden Markov Model (HMM)*, das *Hidden Semi Markov Model*, das *hierarchische HMM*, *Conditional Random Fields* und *Dynamic Time Warping*. Allerdings ist deren Leistung durch die empirisch entworfenen Low-Level-Merkmale begrenzt, weshalb sie im Rahmen dieser Arbeit nicht weiter betrachtet werden.

In den letzten Jahren haben sich auch in diesem Teilbereich der Computer Vision neuronale Netze etabliert, um räumliche und zeitliche Merkmale zur Erkennung von Phasen in OP-Videos automatisiert zu extrahieren. Im Folgenden wird eine Auswahl relevanter Arbeiten vorgestellt.

Twinanda et al. beispielsweise verwenden für ihr *EndoNet* im Wesentlichen eine AlexNet-Architektur zur Extraktion von Merkmalen auf Videoebene. Sie nutzen einen Multi-Task-Ansatz, bei dem sowohl die Erkennung von Instrumenten als auch die Phasenerkennung verwendet wird. Nach der Merkmalsextraktion verläuft die restliche Pipeline mit klassischen Methoden. Die extrahierten Merkmale werden zur Phasenerkennung zunächst an eine *SVM* weitergeleitet. Die Phase wird anschließend durch ein hierarchisches HMM bestimmt (vgl. Abbildung 3.8).

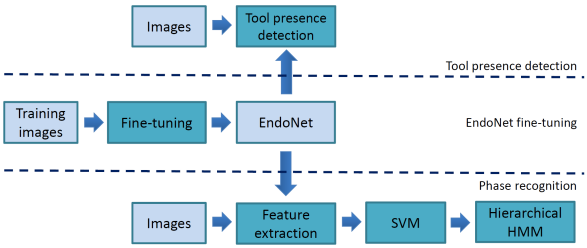


Abbildung 3.8: Komplette Erkennungs-Pipeline nach Twinanda et al. (Quelle: [151])

Die Grundannahme hinter diesem Ansatz basiert auf der erfolgreichen Nutzung von Werkzeugsignalen zur Phasenerkennung in früheren Arbeiten. Eine zweite Annahme der Autoren war, dass durch einen Multitasking-Ansatz mehr diskriminierende Merkmale aus einem Datensatz gewonnen werden können.

Entsprechend ist die Architektur so gewählt, dass eine automatische Erkennung der sichtbaren Werkzeuge in die Phasenerkennung mit einfließt (vgl. Abbildung 3.9) [151].

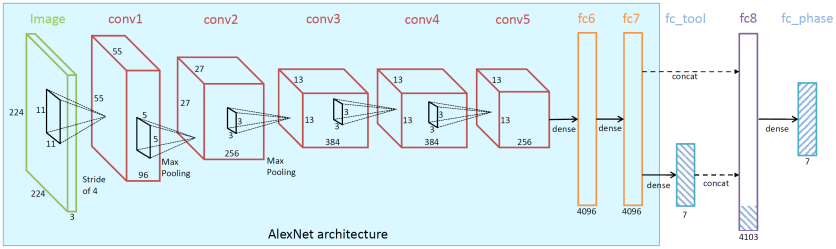


Abbildung 3.9: Schematische Darstellung der EndoNet-Architektur (Quelle: [151])

Zusätzlich zu den Werkzeuginformationen nutzten Nakawala et al. in [115] weitere Anhaltspunkte wie Management-Tools, Ontologie und Produktionsregeln, um die Leistung zu verbessern. Einige Arbeiten extrahierten auch die Optical Flows und nutzten die Bewegungsinformationen, zur Verbesserung des Lernverlaufs von Modellen. Diese Methoden leiden unter den zusätzlichen Kosten für Annotationen bei Multitasking oder verursachen zusätzlichen Rechenaufwand, um andere Modalitäten, z. B. Optical Flows, zu erhalten.

Jin et al. stellten später *SV-RCNet* vor. Dieses integriert ResNet- und LSTM-Module zur Extraktion sequentiell räumlich-zeitlicher Merkmale für die Erkennung chirurgischer Phasen. SV-RCNet wurde außerdem mit einer Post-Processing-Methode namens *Pre-Knowledge Inference (PKI)* erweitert, die alle Phasenvorhersagen der vergangenen Frames über Akkumulatoren für die einzelnen Phasen verwaltet. Über die Anzahl der in jeder Phase klassifizierten Frames wird die Konfidenz der Phasenprädiktion kontinuierlich verbessert [81]. Die SV-RCNet-Architektur wurde in [82] weiterentwickelt, so dass zusätzlich die Präsenz von Instrumenten erkannt und als ergänzendes Merkmal zur Phasenerkennung genutzt werden kann. Die beiden Zweige des Netzwerks zur Instrumenten- bzw. Phasenerkennung beeinflussen sich gegenseitig über eine

sog. *Mapping Matrix*. Die Autoren bezeichnen dieses Netzwerk als *MTRCNet-CL*.

Architekturbedingt und begrenzt durch GPU-Ressourcen können LSTM- und 3D-CNN-Ansätze lediglich kurze zeitliche Abhängigkeiten abbilden. Chirurgische Eingriffe, die von automatisierten Erkennungssystemen profitieren können, dauern üblicherweise allerdings mehrere Minuten bis Stunden. Aufgrund der großen Vielfalt der chirurgischen Szenen in jeder Phase und der häufig vorhandenen Artefakte reicht die Betrachtung von Kurzzeitinformationen für eine genaue Erkennung nicht aus. Um die weitreichende zeitliche Dynamik effektiv zu erfassen, hat die gleiche Forschungsgruppe um Jin et al. zusätzlich das *TMRNet* entwickelt. Diese nutzt zusätzlich zu LSTM-Modulen eine *Memory-Bank*, um Merkmale aus länger andauernden und unterschiedlich skalierten Zeitphasen für die chirurgische Phasenerkennung einfließen zu lassen [83].

Im Gegensatz zu den bisher betrachteten Modellen verwenden Czempiel et al. für *TeCNO* anstelle von LSTM-Modulen erstmals mehrstufige temporale Faltungsnetze (*MS-TCNs*), um längere zeitliche Kontexte erfassen zu können und Videos für die Segmentierung von Aktionen hierarchisch zu verarbeiten. Die Einführung gestapelter Prädiktorenstufen ermöglicht eine schrittweise Verfeinerung der anfänglichen Vorhersagen der vorherigen Stufen [29]. Yi et al. zeigen außerdem in [167], dass durch separates Training der einzelnen Stufen signifikante Verbesserungen in der Erkennungsqualität erreicht werden können.

Gao et al. verwenden in ihrem *Trans-SVNet* ebenfalls ein TCN zur Merkmalsextraktion. Darüber hinaus fusionieren sie die resultierenden zeitlichen und räumlichen Merkmale in einem *Aggregationsmodell*, wodurch, im Vergleich zu anderen Ansätzen, die räumlichen Abhängigkeiten stärker in die Phasenerkennung einfließen. Innerhalb des Aggregationsmodells kommen dafür zwei Transformer-Schichten zum Einsatz [52].

Die zuvor diskutierten Ansätze nutzen allesamt lediglich Informationen auf Frame-Ebene, um auf die aktuelle Phase zu schließen. Ding and Li entwickeln in [38] zusätzlich eine *Segment-Level Semantik*, um Informationen aus den angrenzenden Frames mit in die Entscheidungsfindung einfließen zu lassen und

somit die Prädiktion auf Frame-Level zu verbessern. Die Idee hinter dem sog. *SAHC* (segment-attentive hierarchical consistency network) ist es, hierarchische, semantisch konsistente Segmente auf hoher Ebene zu extrahieren und diese zu verwenden, um die durch mehrdeutige Frames verursachten fehlerhaften Vorhersagen auf niedrigerer Ebene zu verfeinern. Abbildung 3.10 stellt am Beispiel eines Ausschnittes eines Cholezystektomie-Videos den Zusammenhang zwischen Frame- und Segment-Ebene dar. Im oberen Abschnitt ist skizziert, wie mehrdeutige Informationen aus Einzelbildern das Modell zu fehlerhaften Vorhersagen für die chirurgische Phasenerkennung veranlassen würden. Der untere Abschnitt zeigt den Unterschied auf Segment-Ebene. Im Vergleich zu Informationen auf Bildebene können Informationen auf Segment-Ebene mehr Unterscheidungsmerkmale für die chirurgische Phasenerkennung bieten.

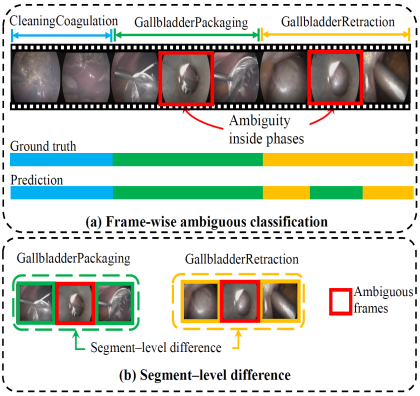


Abbildung 3.10: Grafische Gegenüberstellung von Merkmalen auf Frame- und Segment-Level zur Erläuterung der SAHC-Funktionsweise (Quelle: [38])

Für die Gewinnung der temporalen Merkmale auf Frame-Ebene kommen ebenfalls TCNs zum Einsatz. Die Merkmale auf Segment-Ebene werden anschließend im sog. *SFE (Segment-Level Feature Extractor)* durch Fusion zeitlich zusammenhängender Frames erzeugt. Um die Informationen zwischen den Ebenen auszutauschen und somit die fehleranfälligen Vorhersagen auf Frame-Ebene zu

verbessern, werden Transformer-Schichten für unterschiedliche Skalierungen innerhalb des sog. *SFA-Moduls (Segment-Frame Attention)* eingeführt.

Tabelle 3.1 zeigt die Ergebnisse der vorgestellten Modelle auf dem Cholec80-Datensatz. Hier ist eine deutliche Steigerung der Ergebnisse über die verschiedenen Ansätze hinweg erkennbar.

Tabelle 3.1: Ergebnisse der Phasenerkennung der vorgestellten Modelle auf dem Cholec80-Datensatz

Methode	Genauigkeit	Präzision	Recall
EndoNet [151]	$81,7 \pm 4,2$	$73,7 \pm 16,1$	$79,6 \pm 7,9$
SV-RCNet [81]	$85,3 \pm 7,3$	$80,7 \pm 7,0$	$83,5 \pm 7,5$
MTRCNet-CL [82]	$89,2 \pm 7,6$	$86,9 \pm 4,3$	$88,0 \pm 6,9$
TMRNet [83]	$90,1 \pm 7,6$	$90,3 \pm 3,3$	$89,5 \pm 5,0$
TeCNO [29]	$88,6 \pm 2,7$	$81,6 \pm 4,1$	$85,2 \pm 10,6$
Trans-SVNet [52]	$90,3 \pm 7,1$	$90,7 \pm 5,0$	$88,8 \pm 7,4$
SAHC [38]	$91,8 \pm 8,1$	$90,3 \pm 6,4$	$90,0 \pm 6,4$

Die bisher vorgestellten Methoden nutzten zum Training und zur Evaluation allesamt Cholec80-Daten. Im Rahmen der EndoVis-Challenge 2019 wurde das HeiChole-Datenset veröffentlicht und in [159] die dort eingesetzten Methoden rudimentär erläutert. Implementierungsdetails sind in der Veröffentlichung nicht dargestellt, sondern lediglich die grobe Methodik skizziert. Die unterschiedlichen Gruppen nutzten im Wesentlichen ähnliche, weiterentwickelte Ansätze, wie in den zuvor beschriebenen Modellen. Beliebt waren u. A. Multi-Tasking, temporale Komponenten wie bspw. LSTMs und Postprocessing mittels Prozesswissen. Die Ergebnisse sind in Tabelle 3.2 dargestellt, während weitere Details zu den einzelnen Phasen Tabelle 3.3 entnommen werden können. Die beiden Ergebnisse des Teams *CAMMA* wurden dabei nachträglich eingereicht, weshalb sie nicht am Wettbewerb teilnehmen konnten. Die Resultate wurden dennoch im Sinne der Vergleichbarkeit veröffentlicht.

Tabelle 3.2: Ergebnisse der Phasenerkennung der teilnehmenden Teams auf dem HeiChole-Datensatz (Quelle: [159]).

Team	durchschn. F1
CAMMA (vortrainiert)	68,78
HIKVision	65,38
CUHK	64,98
CAMMA (nicht vortrainiert)	63,60
MEVIS	57,30
NCT Dresden	49,00
Wintegral	42,47
CAMI-SIAT	38,65
VIE-PKU	33,29
IGI Medical Technologies	23,93

Tabelle 3.3: Detailliertes Ergebnis der einzelnen Phasen des Gewinnerteams und Durchschnitt über alle Teams beim HeiChole-Benchmark 2019 (Quelle: [159]).

Team	F1-Score						
	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
HIKVision	88,49	86,50	46,92	74,88	45,90	66,19	62,03
Alle	77,59	73,97	36,73	67,29	42,76	47,65	41,54

Zusammenfassend ist festzuhalten, dass in bisherigen Deep Learning-basierten Arbeiten zur Workflowanalyse in OPs unterschiedliche Netzwerkarchitekturen, wie LSTMs, 3D-CNNs, TCNs oder Transformer, zum Einsatz kamen und mittels verschiedener Ansätze weiter optimiert wurden. Hierbei wurden u. A. Multi-Tasking, mehrstufige Auswerteverfahren oder Informationsaggregation verwendet. Die dargelegten Lösungsansätze betrachten dabei nahezu alle lediglich eine einzige Informationsquelle, im Regelfall das Endoskopbild, als Eingabedaten. Der Mehrwert von diversen Informationsquellen zur Gewinnung von weiteren und insgesamt besser diskriminierenden Merkmalen wird kaum diskutiert. Weiterhin werden zwar über diverse Ansätze temporale Abhängigkeiten

abgebildet, die eigenen prädizierten OP-Phasen der jeweiligen Modelle werden aber nicht explizit weiter berücksichtigt. Über diese Information kann aber nicht nur das Auftreten von Phasen, sondern auch auf deren Dauer geschlossen werden. Da die Anzahl der Phasenübergänge im Vergleich zur jeweiligen Phasendauer nur sehr gering ausfällt, kann die Berücksichtigung der bisher erkannten Phasen zur Verbesserung der aktuellen Prädiktion beitragen. Außerdem fällt auf, dass die Entwicklung der unterschiedlichen Arbeitsgruppen zu immer komplexeren Modellen führt.

Die diskutierten Modelle befassen sich fast alle in erster Linie mit der OP-Phasenerkennung und nicht mit dem Workflow innerhalb der OP. Dies wird Lediglich in [115] thematisiert. Das ist auch die einzige Arbeit, die neben dem Instrumenteneinsatz weitere Informationen in den Entscheidungsprozess einbezieht und damit erst eine fundierte Ablaufanalyse ermöglicht. Das in der vorliegenden Arbeit dargestellte System kann durch die Unterteilung in verschiedene Teilsysteme mit unterschiedlichen Erkennungsschwerpunkten und seine leichte Erweiterbarkeit um weitere Teilsysteme ebenfalls einen Beitrag zur Workflowanalyse leisten und ist somit den reinen Phasenerkennungssystemen überlegen.

3.4 Datensätze für datengetriebene Algorithmenentwicklung

Die Grundlage für datengetriebene Entwicklungsmethoden (vgl. Abschnitt 2.2.3) bilden sehr große und möglichst diverse Datenbestände aus der jeweiligen Anwendungsdomäne. Neben den zu untersuchenden Daten werden für das Modelltraining zusätzlich Annotationen mit relevanten Informationen, bspw. an welcher Stelle eines Bildes sich das gesuchte Objekt befindet, benötigt. Hierfür ist weitreichendes Domänenwissen notwendig, das üblicherweise nur Experten auf diesem Gebiet leisten können. Hinzu kommt, dass die Hardware-Anforderungen für das Training komplexer Modelle mit deren Größe steigen, was wiederum

die Gesamtkosten des Trainings stark erhöht und oft nur von großen Unternehmen getragen werden kann. Aus diesem Grund kommen häufig vortrainierte Modelle zum Einsatz, die mit Hilfe von Transfer-Learning feinabgestimmt (engl. *finetuning*) werden. Dennoch ist eine Reihe von gelabelten Trainingsdaten aus der Anwendungsdomäne nötig. Die Erzeugung neuer Datensätze ist extrem zeit- und ressourcenaufwändig und bedarf großer Sorgfalt, um die Datenqualität ausreichend hoch zu halten. Aus diesem Grund werden, insb. für das Finetuning vortrainierter Modelle, oftmals öffentlich verfügbare Datensätze verwendet.

In den folgenden Abschnitten werden öffentlich verfügbare Datensätze für die in dieser Arbeit relevanten Anwendungsfälle *Personen-Re-Identifikation* und *Analyse medizinischer Arbeitsabläufe* bzw. *Analyse endoskopischer Videos* im Speziellen diskutiert.

In Bezug auf Videomaterial mit medizinischem Kontext, das für diese Arbeit genutzt werden könnte, sind zum Zeitpunkt der Erarbeitung nur wenige Datensets öffentlich verfügbar. Dies ist zum einen begründet durch die Herausforderungen bzgl. Datenschutz, da es sich im medizinischen Umfeld und vor allem die Patientinnen und Patienten betreffend meist um hoch sensible Daten handelt. Zum anderen aber auch dadurch, dass Fragestellungen in diesem Bereich häufig sehr spezifisch sind und die verfügbaren Daten entsprechend auf diese spezifischen Eigenschaften fokussiert sind. Dies erschwert eine Nutzung für andere Anwendungsfälle. Das folgende Unterkapitel stellt eine Auswahl der relevantesten Rechercheergebnisse dar.

3.4.1 Datensätze für Raumbeobachtung und Personenerkennung

Im weiteren Verlauf werden zunächst Datensets für die Beobachtung von Abläufen und Bewegungen innerhalb eines Raumes oder einer bestimmten Fläche sowie für die Erkennung von Personen und deren Pose betrachtet. Die Erkenntnisse fließen in die Arbeiten der Kapitel 5 und 6 ein.

Ein Datenset, welches speziell für das Testen von Methoden zur Posenerkennung mit medizinischen Aufnahmen angefertigt wurde, ist das *MVOR-Datenset* der Universität Straßburg. Es besteht aus 732 synchronisierten Einzelbildern aus mehreren Perspektiven, welche mit insgesamt drei RGB-D Kameras während realer klinischer Eingriffe aufgezeichnet wurden. Die Daten enthalten neben den Farb- und Tiefenbildern zusätzlich Parameter für die Kamerakalibrierung, Bounding Boxen für die sichtbaren Personen sowie Annotationen für 2D- und 3D-Posen [143]. Abbildung 3.11 zeigt eine Auswahl von Bildern aus dem Datensatz. Hier ist auch zu erkennen, dass die Personen teilweise farbige Haarnetze tragen und dadurch eine Unterscheidung zulassen. Eine eindeutige Annotation der Individuen, wie es für eine Re-Identifikation notwendig wäre, ist allerdings nicht vorhanden. Da der öffentliche Teil des MVOR-Datensets außerdem kein Videomaterial beinhaltet, sondern lediglich aus einer Sammlung von Einzelbildern besteht und damit keine Bewegungsanalyse ermöglicht, ist es für eine weitere Verarbeitung im Sinne der geplanten Konzepte nicht geeignet. Laut den Erstellern des Datensets ist zwar eine Veröffentlichung der zugehörigen Videos geplant, zum Zeitpunkt der Fertigstellung dieser Arbeit waren diese jedoch noch nicht verfügbar. Dadurch ist der MVOR-Datensatz nicht zur Untersuchung der Konzepte dieser Arbeit (vgl. Kapitel 4) geeignet.

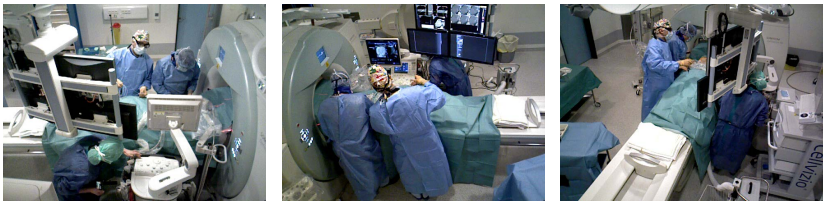


Abbildung 3.11: Beispielbilder des MVOR-Datasets (Quelle: [143])

Eine weitere Quelle für Videomaterial sind Aufnahmen aus sehr ähnlichen Domänen, wie bspw. der *Veterinärmedizin*. Diese besitzen ähnliche Eigenschaften bezüglich der Perspektive und der Bewegungsgeschwindigkeit von Personen und können deshalb für erste Tests verwendet werden. Da nicht mit menschlichen Patienten gearbeitet wird, sind die Aufnahmen weniger kritisch und deshalb

auch einfacher zu erhalten. Es ist zwar kein Datensatz mit einer Vielzahl solcher Videos bekannt, aber Einzelvideos sind über gängige Online-Videoportale auch frei zugänglich verfügbar. Auszüge einer solchen Aufnahme sind in Abbildung 3.12 dargestellt [30].

Abbildung 3.12: Aufnahmen einer Deckenkamera aus einer Tierarztpraxis (Quelle: [30])

Die *Re-Identifikation von Personen* ist ein aktives Forschungsfeld und entsprechend sind diverse spezielle Datensets für diesen Anwendungsfall verfügbar. Allerdings finden sie zumeist im öffentlichen Raum und mit Personen in Alltagskleidung statt. Ein Datenset mit Personenzuordnung im medizinischen Kontext ist aktuell nicht bekannt. Tabelle 3.4 zeigt eine Übersicht über häufig verwendete Datensätze für das Training und die Validierung von Re-ID-Modellen ohne medizinischen Bezug. Aufgrund des Umfangs und der Verbreitung innerhalb der Re-ID-Community wird im Verlauf der Arbeit lediglich der *Market1501*-Datensatz weiter betrachtet (vgl. Abschnitt 5.3). Dieser wurde von sechs Kameras aus verschiedenen Perspektiven vor einem Campus-Supermarkt aufgenommen. Entsprechend enthalten die Bilder in der Regel Personen in ihrer normalen Alltagskleidung. Der Datensatz enthält insgesamt 32.668 Bounding Boxes von 1.501 verschiedenen Identitäten [171].

3.4.2 Laparoskopische Datensets

Dieser Abschnitt beschreibt Datensets, die während laparoskopischer Eingriffe aufgezeichnet und speziell für die Erkennung von Merkmalen solcher Operationen erstellt wurden. Diese werden für die Arbeiten in Kapitel 7 benötigt.

Tabelle 3.4: Übersicht von häufig verwendeten Datensätzen für das Training und die Validierung von Re-Identifikationsmodellen nach [175].

Datensatz	# IDs	# Bilder	# Kameras
Market1501 [171]	1501	32668	6
CUHK03 [105]	1467	28192	2
DukeMTMC [133], [172]	1812	36411	8
MSMT17 [163]	4101	126411	15
VIPeR [61]	632	1264	2
GRID [109]	251	1275	6
CUHK01 [104]	971	3882	2

Der *Cholec80*-Datensatz enthält 80 intraoperative laparoskopische Videos von Cholezystektomie-Eingriffen, die von 13 Chirurgen durchgeführt wurden. Die Videos wurden mit einer Auflösung von 1920×1080 oder 854×480 mit 25 Frames pro Sekunde (FPS) aufgenommen. Jedes Einzelbild ist mit einer von insgesamt sieben chirurgischen Phasen annotiert, welche von einem leitenden Chirurgen eines Partnerkrankenhauses der Autoren festgelegt wurden. Zusätzlich sind sieben verschiedene Instrumenten-Klassen enthalten, die mit 1 FPS die Präsenz des jeweiligen Instruments in einem Frame annotieren. In Abbildung 3.13 sind die Instrumente zu den Klassen dargestellt [151].

(a) Grasper **(b)** Bipolar **(c)** Hook **(d)** Clipper

(e) Scissors **(f)** Irrigator **(g)** Specimen Bag

Abbildung 3.13: Darstellung der Instrumentenklassen im Cholec80-Datensatz (Quelle: [151])

Die Cholec80-Videos stammen zwar von unterschiedlichen OP-Teams, die aber alle in der gleichen Einrichtung tätig sind. Dies führt u. A. dazu, dass die Varianz im Ablauf der einzelnen Eingriffe sehr gering und daher wenig geeignet als allgemeingültige Repräsentation dieser OP-Art ist [111]. Dennoch hat Cholec80 eine große Bedeutung im Forschungsumfeld, da nach wie vor viele Arbeiten im Bereich der OP-Phasenerkennung zumindest teilweise auf diesem Datenset basieren und entsprechend eine Vergleichbarkeit hergestellt werden kann (vgl. Abschnitt 3.3).

Bei *HeiChole* handelt es sich um einen Datensatz mit insgesamt 33 laparoskopischen Cholezystektomie-Videos aus drei chirurgischen Zentren mit einer Gesamtoperationszeit von 22 Stunden. Davon wurden jedoch nur 24 Videos aus dem Trainingsset zur öffentlichen Verfügung bereitgestellt. Außerdem gilt es zu beachten, dass nur zwei der drei Einrichtungen in den Videos vertreten sind. Die restlichen neun Videos wurden für die Evaluation im Rahmen der *Surgical Workflow and Skill Analysis Challenge*, einer Sub-Challenge der *2019 International Endoscopic Vision Challenge (EndoVis)*, zurückgehalten. Die Einzelbilder sind, analog zu Cholec80, mit sieben chirurgischen Phasen annotiert. Außerdem sind im gesamten Datenset 250 Phasenübergänge, 5.514 Vorkommen von vier chirurgischen Handlungen, 6.980 Vorkommen von 21 chirurgischen Instrumenten aus sieben Instrumentenkategorien sowie 495 Fertigkeitsklassifizierungen in fünf Fertigkeitsdimensionen in den Annotationen enthalten. In den verschiedenen Einrichtungen wurden Instrumente unterschiedlicher Hersteller eingesetzt, was auch zu verschiedenen Eigenschaften der Daten führte. Dadurch entstanden Aufnahmen mit einer Auflösung von 960×560 Pixeln mit 25 FPS, 1920×1080 Pixeln mit 50 FPS sowie 720×576 Pixeln mit 25 FPS. Die Instrumentenlabels sind, auch aufgrund der höheren Variabilität durch die verschiedenen Hersteller, nicht komplett identisch zu den Labels aus Cholec80, sodass hier keine genaue Vergleichbarkeit hergestellt werden kann [159].

HeiChole bietet im Vergleich zu Cholec80 mehr Variabilität in der Auflösung, Framerate, Videolänge, Instrumentennutzung und in den Ablaufsequenzen und spiegelt folglich wesentlich besser den realen Operationsalltag wider. Dadurch,

dass das Datenset neuer ist als Cholec80, gibt es allerdings bisher merklich weniger Arbeiten, die Erkennungswerte zum Vergleich liefern.

3.5 Fazit & Abgrenzung

Der Stand der Wissenschaft und Technik zeigt, dass bzgl. der Methodik zur Erfassung von Objekten, Personen und deren Bewegungen und Handlungen sowie der Analyse bestimmter Prozessabläufe sehr rege Forschungsaktivitäten stattfinden und für genannte Teilaspekte bereits robuste Modelle existieren. Einer der Hauptgründe, warum diese Resultate für medizinische Anwendungen bisher nicht im Klinikalltag ankommen, ist der Mangel an stabilen und intuitiven Systemen der Medizintechnikhersteller zur Aufzeichnung konsistenter Trainingsdaten in ausreichender Menge.

Die Endoscopic Vision Challenge *EndoVis*, die regelmäßig während der internationalen Konferenz der Medical Image Computing and Computer Assisted Intervention Society (MICCAI) stattfindet, ist die größte Quelle für SDS-Datensammlungen. Sie besteht aus mehreren jährlichen Unterwettbewerben, die die Verfügbarkeit neuer öffentlicher Datensätze für die Entwicklung und das Benchmarking von Methoden unterstützen. Generell ist jedoch die Qualitätskontrolle bei biomedizinischen Herausforderungen und die gemeinsame Nutzung von Daten immer noch ein Problem [111]. Maier-Hein et al. listen in ihrer Arbeit weitere Datensets, die hier nicht diskutiert wurden. Die aufgezeigte Auswahl erfolgte aufgrund der zum Erarbeitungszeitraum höchsten Übereinstimmung mit den vorliegenden Anforderungen bzgl. Menge, Qualität und Vergleichbarkeit der Daten bzw. deren Ergebnisse.

Die Erläuterungen in den vorangegangenen Abschnitten zeigt aber auch, dass bisher kein konsistentes Datenset für alle Teilaspekte der vorliegenden Arbeit existiert. Insb. fehlen zusammenhängende Daten, die chirurgische Eingriffe mit verschiedenen Kamerasystemen aus unterschiedlichen Perspektiven zeigen. Zur Detektion und Bewegungsanalyse von Personen existieren lediglich Daten aus

anderen Domänen und für die Instrumentenerkennung aus der Tisch-Perspektive sind gar keine passenden Daten bekannt. Außerdem fehlt die Kombination mit zusätzlichen (Sensor-)Informationen, bspw. aus Mikrofonen, Beschleunigungssensorik oder dem KIS, was nicht der Hauptfokus dieser Arbeit bildet, aber konzeptionell interessant wäre. Bestehende Datensets sind im Regelfall sehr spezifisch auf konkrete Fragestellungen ausgelegt und erlauben aufgrund der dadurch vorhandenen Informationen und Annotationen nur bedingt den Einsatz für angrenzende Anwendungsfälle.

Als Konsequenz daraus lässt sich ableiten, dass entweder mit nicht exakt passenden Daten gearbeitet werden muss (vgl. Teile von Kapitel 5), was zu nicht optimalen Ergebnissen führen kann, oder eigene Daten aufgezeichnet werden müssen (vgl. Abschnitt 5.3 & Kapitel 6), was wiederum extrem zeit- und kostenintensiv ist.

Zur Abgrenzung vom aktuellen Stand der Technik sollen zunächst die einzelnen Teilsysteme dieser Arbeit betrachtet werden.

In Bezug auf die Posenerkennung, Bewegungsanalyse und Re-ID wird im Gegensatz zu den üblichen bekannten Forschungsarbeiten im Rahmen dieser Arbeit der *Einfluss von medizinischer Kleidung* und den damit einhergehenden spezifischen Herausforderungen genauer untersucht und Lösungsmöglichkeiten erarbeitet.

Der Trend der vorgestellten Modelle zur OP-Phasenerkennung tendiert dazu, dass diesen zur Performance-Steigerung immer komplexere Architekturen zugrundeliegen. Damit geht üblicherweise auch ein höherer Ressourcenbedarf sowohl an die benötigte Hardware als auch an die Quantität und Qualität der benötigten Trainingsdaten einher. Aus diesem Grund setzt die vorliegende Arbeit eine OP-Phasenerkennung um, die anstelle von komplexen Bewegungsanalysen im Endoskopvideo auf Frame-Ebene, lediglich auf den *zeitlichen Abfolgen der eingesetzten Instrumente* basiert (vgl. Abschnitte 6 & 7). Darüber hinaus soll die Phasenprädiktion durch den *Einbezug der vorangegangenen Modellausgaben* weiter optimiert werden (vgl. Abschnitt 7).

Auf übergeordneter Ebene betrachten die in diesem Kapitel dargelegten Lösungsansätze im Regelfall ausschließlich ein Teilproblem der komplexen OP-Workflowanalyse. Abgrenzend dazu wird in dieser Arbeit ein Konzept erarbeitet, welches es erlaubt, Teilsysteme für einzelne Lösungen individuell zu entwickeln und anschließend zu einem komplexen Gesamtsystem zu konsolidieren, sodass durch *Multi-Tasking-Ansätze* zusätzliche Erkenntnisse über den Operationsverlauf gewonnen werden können. Dabei soll es möglich sein *mehrere Eingabequellen*, wie bspw. Videos aus Laparoskopen, aus Deckenkameras und Kameras am Instrumententisch, aber auch weitere Informationsquellen wie Mikrofone oder KIS-Daten gleichermaßen zu nutzen (vgl. Abschnitt 4.3). Durch die hohe Modularität ist eine *Erweiterbarkeit* durch weitere Teilsysteme mit geringem Aufwand möglich.

4 Konzeption eines Systems zur Erfassung von Kontexten laufender Operationen zur Deduktion von OP-Phasen

Für die Konzeptionierung eines Erkennungssystems zur automatischen Phasenerkennung der laparoskopischen Cholezystektomie ist in erster Linie ein ausgeprägtes Verständnis über den typischen Ablauf der OP und der zugehörigen Prozesse sowie den Personalaufwand und die Aufgabenverteilung von Bedeutung. In dieser Hinsicht werden in den folgenden Abschnitten die Erkenntnisse aus den Kapiteln 2 und 3 tiefergehend analysiert und modelliert. Hierfür werden zunächst allgemeine Anforderungen an ein solches Erkennungssystem erhoben. Anschließend werden technische Modelle aus den Grundlagen abgeleitet, die die Basis für die Auswahl geeigneter Sensorik und Algorithmen bilden. Daraus können schließlich spezifischere Anforderungen definiert und das Gesamtkonzept für die Komponenten und die Architektur des Erkennungssystems formuliert werden. Zusätzlich wird noch eine Risikobetrachtung des vorliegenden Konzepts durchgeführt.

4.1 Anforderungen an das Erkennungssystem

Zur Erarbeitung eines Konzepts für das Gesamtsystem werden zunächst Anforderungen definiert. Diese basieren auf den Erkenntnissen aus den Kapiteln 2 und

3 sowie weiteren Rahmenbedingungen, die bspw. direkt aus dem OP-Umfeld entstammen. Folgende Anforderungen wurden dabei festgelegt:

[*AF-01*] Das Erkennungssystem soll eine automatisierte Erkennung einzelner Abschnitte (OP-Phasen) der laparoskopischen Cholezystektomie ermöglichen.

[*AF-02*] Das System soll als onlinefähiges System eine Auswertung während der OP-Laufzeit inklusive Live-Feedback ermöglichen.

[*AF-03*] Das System soll Verzögerungen im OP-Ablauf erkennen können.

[*AF-04*] Das System muss Personen innerhalb des OP-Saals erkennen können.

[*AF-05*] Das System soll Personen innerhalb des OP-Saals tracken können.

[*AF-06*] Das System muss die OP-Instrumente, die bei der laparoskopischen Cholezystektomie zum Einsatz kommen, erkennen können.

[*AF-07*] Das System muss den Zeitpunkt der Nutzung eines OP-Instruments erkennen können.

[*AF-08*] Das System soll Informationen über die Zustände von den Geräten im OP-Saal auslesen können.

[*AF-09*] Das System soll minimalen Installationsaufwand im OP-Saal verursachen.

[*AF-10*] Das System soll bei der Nutzung keinen Mehraufwand für das OP-Team verursachen.

[*AF-11*] Das System muss bei Beleuchtungsstärken ≥ 4 Lux funktionieren.

[*AF-12*] Das System soll lediglich ambiente Sensorik verwenden, um Probleme mit Hygieneanforderungen im Operationssaal durch zusätzliche körpernahe Sensorik zu vermeiden.

[*AF-13*] Das System soll unabhängig von Medizintechnikgeräte-Herstellern funktionieren.

4.2 Technische Analyse der laparoskopischen Cholezystektomie

Um die Konzeptionierung eines technischen Systems zur automatisierten Erkennung von Prozessen und Phasen im OP zu ermöglichen, werden in diesem Abschnitt die Erkenntnisse aus Kapitel 2 und insbesondere Abschnitt 2.1 tiefergehend analysiert. Eine technische Modellierung der Analysen mittels Unified Modeling Language (UML) ermöglicht die formelle grafische Darstellung der relevanten Zusammenhänge und eine anschließende strukturierte Erhebung weiterer Anforderungen an das zu entwickelnde System.

4.2.1 Modellierung der laparoskopischen Cholezystektomie

Zunächst werden die beteiligten Akteure und deren Aufgaben in der laparoskopischen Cholezystektomie, wie in Abschnitt 2.1.5 diskutiert, genauer betrachtet. Abbildung 4.1 skizziert das daraus resultierende Use-Case-Diagramm. Hierbei sind prä-, intra- und postoperative Prozesse berücksichtigt. Die Einteilung in die Phasen wird gemäß der Spezifikation von Use-Case-Diagrammen hier nicht ersichtlich, kann aber aus den nachfolgenden Aktivitätsdiagrammen gelesen werden.

Für die Konzeption des Erkennungssystems ist neben den Akteuren und Prozessen auch die Abfolge von Handlungen und Abläufen relevant. Das Aktivitätsdiagramm in Abbildung 4.2 zeigt dies für den gesamten OP-Ablauf. Dabei sind die einzelnen Prozesse zusätzlich, dargestellt über drei Spalten, in die Phasen des perioperativen Prozessablaufs prä-, intra- und postoperativ eingeordnet.

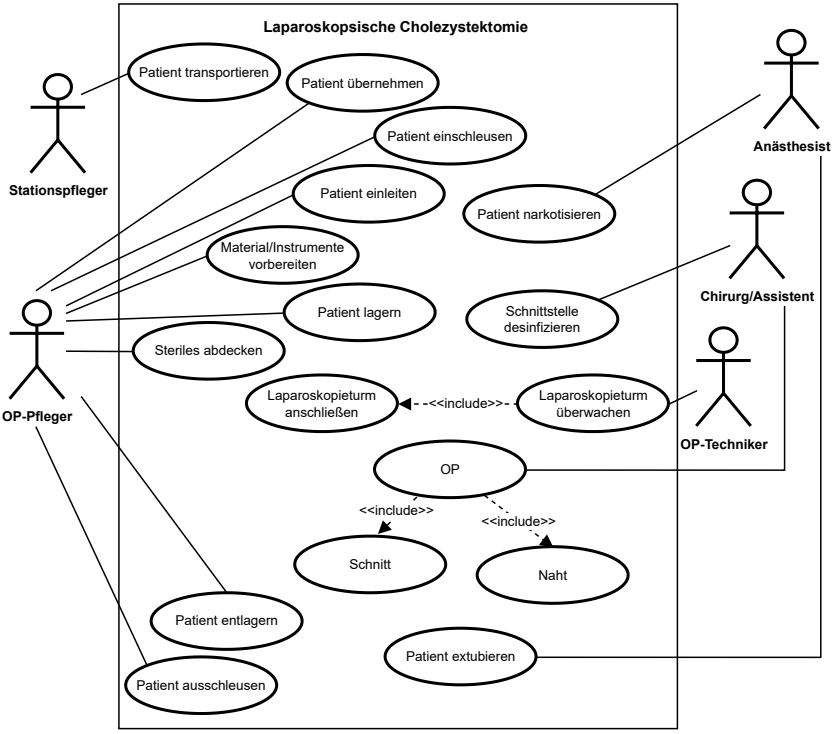


Abbildung 4.1: Use-Case-Diagramm der laparoskopischen Cholezystektomie zur Modellierung der beteiligten Akteure und deren Aufgaben.

Für die Erfassung einzelner Prozesse ist es zudem relevant, ob diese innerhalb oder außerhalb des OP-Saals stattfinden: Innerhalb eines hochgradig integrierten OP-Saals ist eine große Menge an Informationen, wie z. B. Gerätenutzung oder Patienteninformationen, implizit verfügbar, außerhalb kann jedoch auf wenig bis gar keine Informationsquellen zugegriffen werden. Das Diagramm aus Abbildung 4.2 wird entsprechend angepasst, um diese Faktoren sichtbar zu machen (siehe Abb. 4.3 bis 4.5).

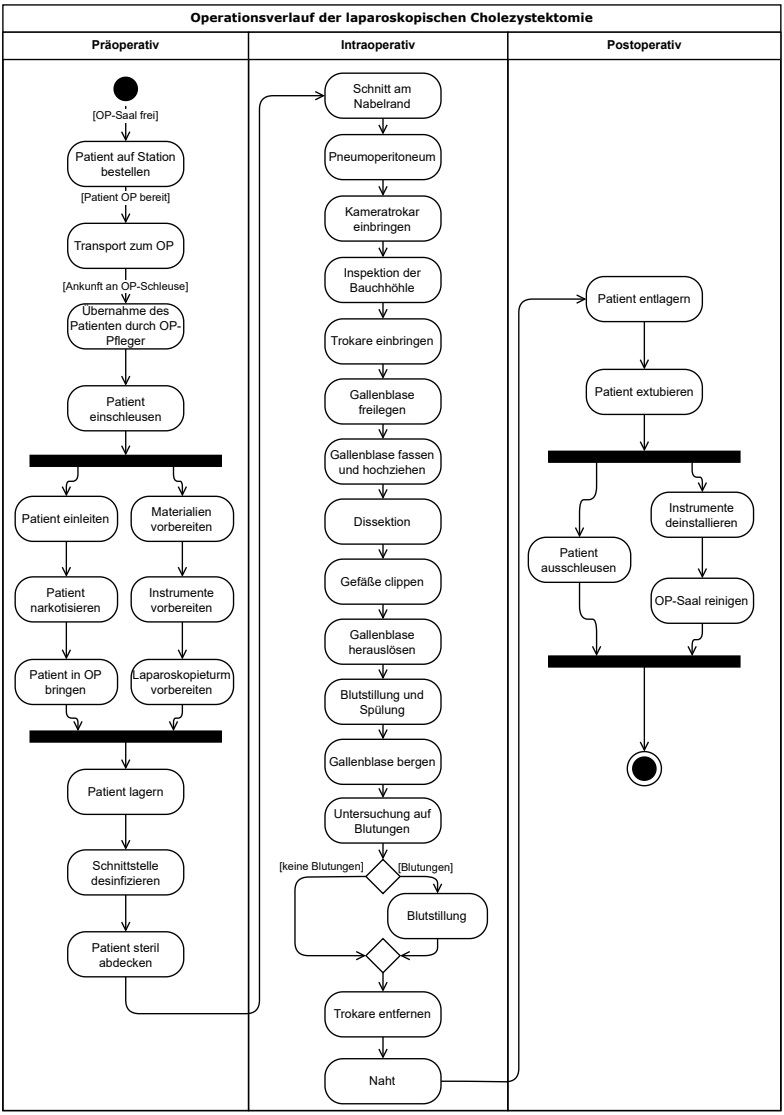


Abbildung 4.2: Aktivitätsdiagramm zum gesamten Ablauf der laparoskopischen Cholezystektomie mit Phaseneinteilung gemäß perioperativem Gesamtprozess.

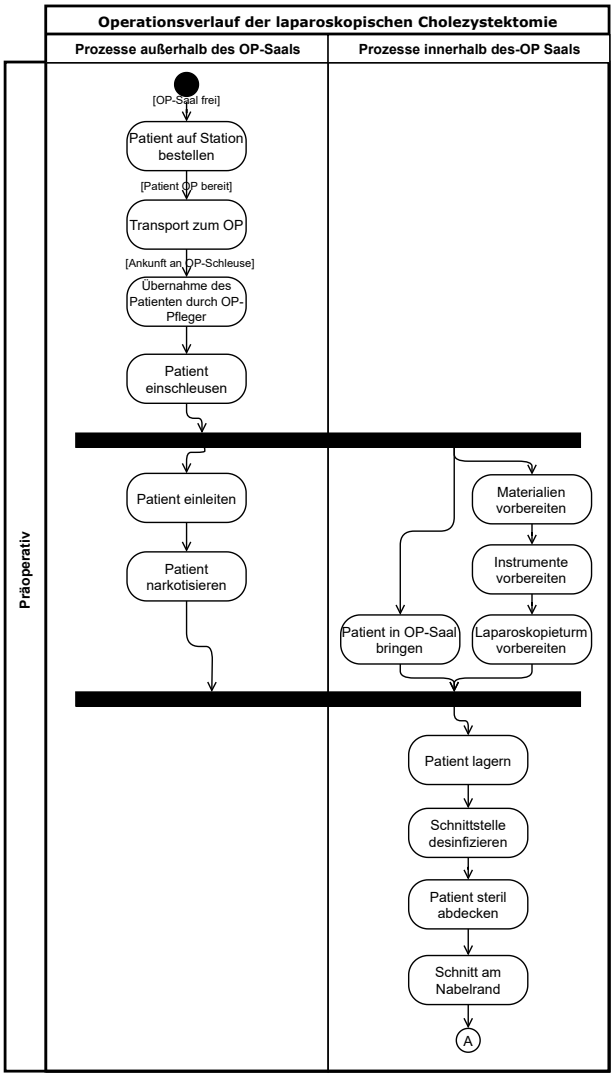


Abbildung 4.3: Aktivitätsdiagramm zum präoperativen Ablauf der laparoskopischen Cholezystektomie mit Einteilung der einzelnen Prozesse in innerhalb und außerhalb des OP-Saals.

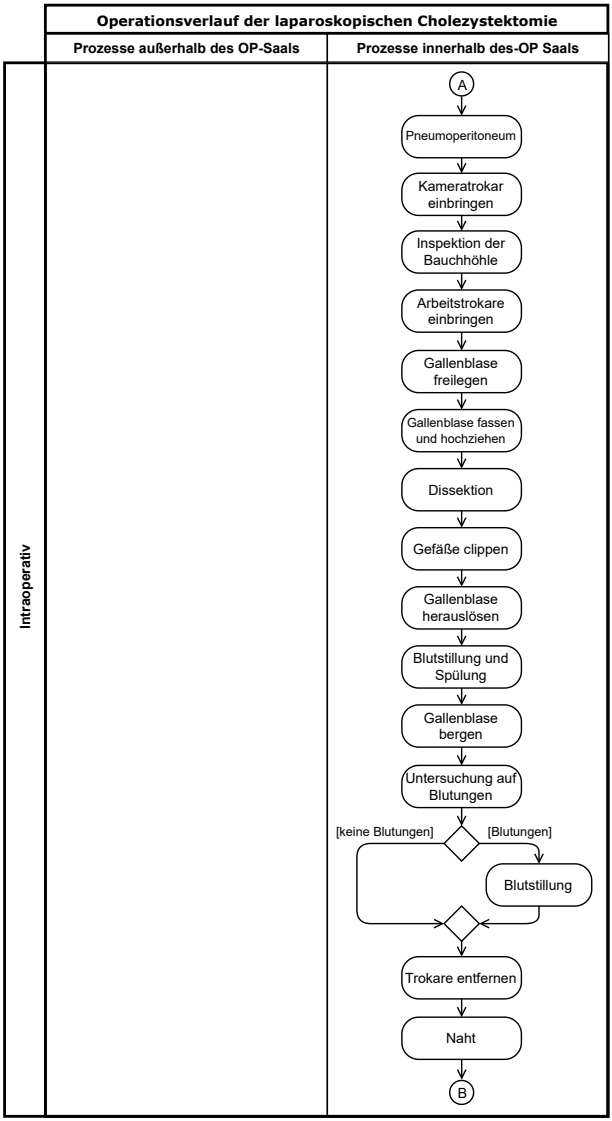


Abbildung 4.4: Aktivitätsdiagramm zum intraoperativen Ablauf der laparoskopischen Cholezystektomie mit Einteilung der einzelnen Prozesse in innerhalb und außerhalb des OP-Saals.

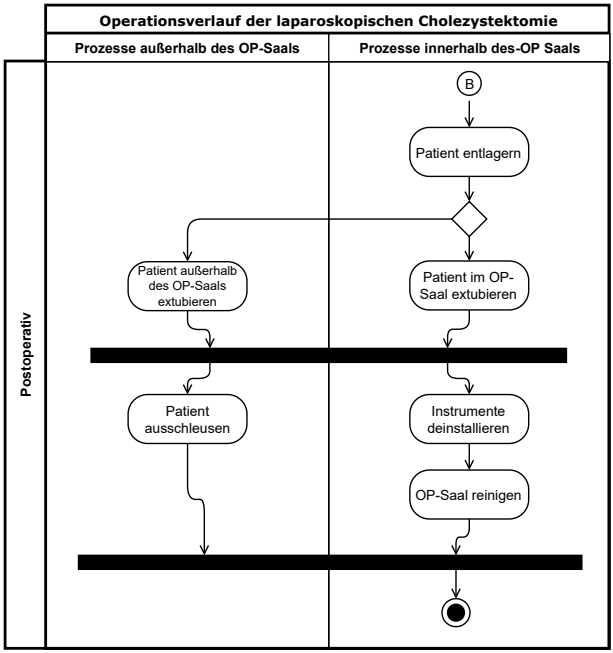


Abbildung 4.5: Aktivitätsdiagramm zum postoperativen Ablauf der laparoskopischen Cholezystektomie mit Einteilung der einzelnen Prozesse in innerhalb und außerhalb des OP-Saals.

Da, wie zuvor erwähnt, innerhalb des Operationssaals bereits zahlreiche relevante Informationen gewonnen werden können und eine kontrolliertere und entsprechend einfacher zu beobachtende Umgebung vorherrscht, werden diese Abläufe in Abbildung 4.6 detaillierter analysiert. Dabei werden die einzelnen OP-Phasen spezifischer und ausführlicher dargestellt. Zugrunde liegen hierfür insbesondere die Erläuterungen aus [25].

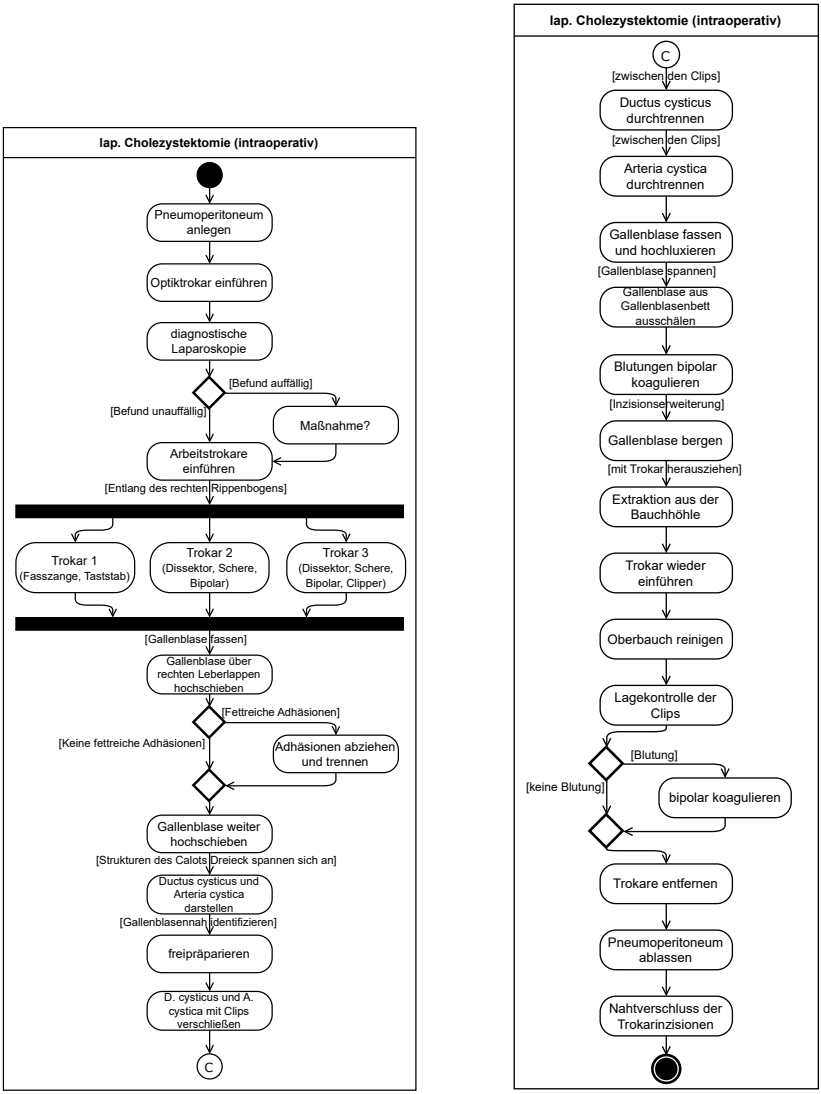


Abbildung 4.6: Detailliertes Aktivitätsdiagramm zum intraoperativen Verlauf der laparoskopischen Cholezystektomie nach [25].

Abbildung 4.7 stellt, in Anlehnung an [94], zusätzlich zum Ablauf der intraoperativen Prozessschritte noch dar, welche OP-Instrumente jeweils zum Einsatz kommen. Dies erlaubt einen Überblick über die verschiedenen Instrumentenwechsel und deren Zusammenhang zu den einzelnen Prozessschritten, was für die weiteren Überlegungen bzgl. automatisierter Erkennbarkeit von OP-Phasen notwendig ist. Neben der theoretischen Betrachtung aus Literaturangaben bietet die Analyse der beiden öffentlich verfügbaren Datensets Cholec80 [151] und HeiChole [159] die Möglichkeit, die Verteilung in realen OPs zu analysieren. Die Konfusionsmatrizen in Abbildung 4.8 zeigen, welche Instrumente in beiden Datensets gleichzeitig im Einsatz sind und in welchen Phasen die jeweiligen Instrumente genutzt werden. Diese Analyse bestätigt weitestgehend die zuvor getroffene Annahme. Es wird aber auch deutlich, dass in der Realität eine größere Unsicherheit in der Verteilung vorherrscht, was allerdings auch zu erwarten war.

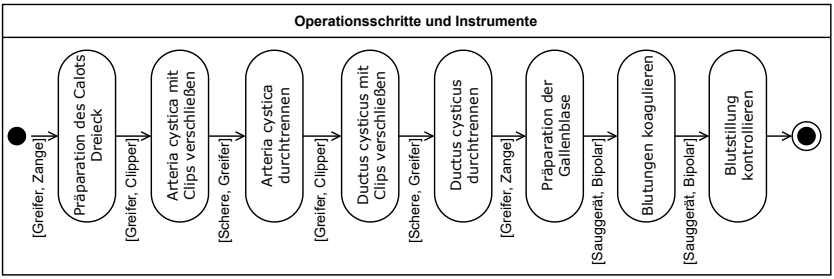
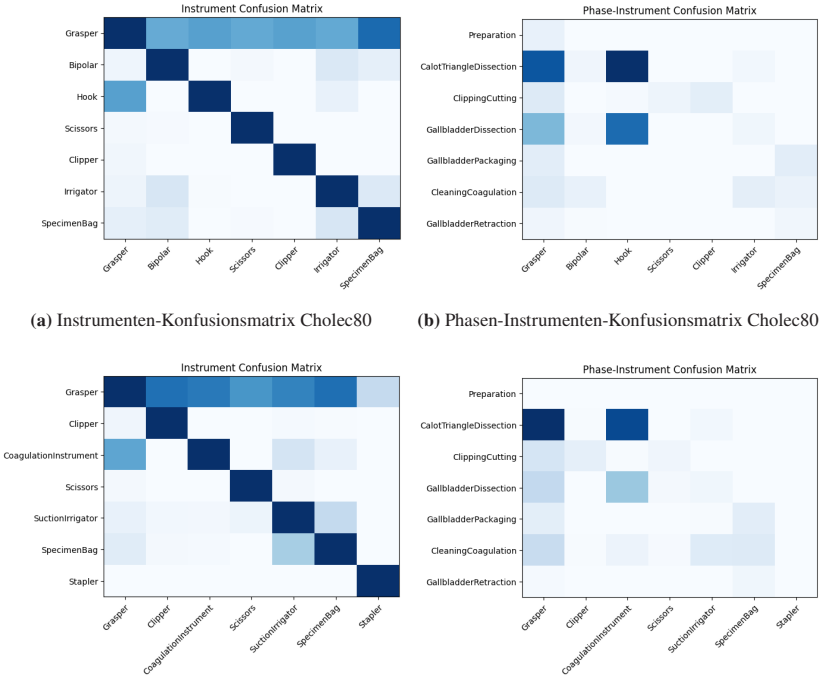


Abbildung 4.7: Aktivitätsdiagramm zum intraoperativen Verlauf der laparoskopischen Cholezystektomie mit Darstellung der jeweils genutzten OP-Instrumente nach [94].

Basierend auf den bisherigen Erkenntnissen lässt sich die laparoskopische Cholezystektomie innerhalb des OP-Saals folglich in insgesamt 28 Operationsschritte gliedern, wie sie in Tabelle 4.1 chronologisch dargestellt sind. Die Tabelle zeigt zusätzlich deren Zuordnung zur jeweiligen Phase im perioperativen Prozess. Insbesondere die 21 Einzelschritte im intraoperativen Abschnitt lassen sich schließlich in acht prägnante, aber differenzierte OP-Phasen gliedern, sodass, zusammen mit den prä- und postoperativen Phasen, insgesamt zehn Phasen den



(a) Instrumenten-Konfusionsmatrix Cholec80 **(b)** Phasen-Instrumenten-Konfusionsmatrix Cholec80

(c) Instrumenten-Konfusionsmatrix HeiChole **(d)** Phasen-Instrumenten-Konfusionsmatrix HeiChole

Abbildung 4.8: Überblick über das Nutzungsverhalten der Instrumente in den Datensets Cholec80 (a, b) und HeiChole (c, d)

kompletten perioperativen Prozess beschreiben (vgl. Tabelle 4.1). Die intraoperative Phasendefinition entspricht im Wesentlichen der Phaseneinteilung der beiden öffentlichen Datensets Cholec80 [151] und HeiChole [159], die in vielen Veröffentlichungen und öffentlichen Challenges zum Thema Workflow-Analyse im OP-Umfeld verwendet werden (vgl. Kapitel 3). Lediglich die Nummerierung unterscheidet sich durch die Hinzunahme der zusätzlichen Phasen. Somit sind die Ergebnisse und Erkenntnisse dieser Arbeit mit dem aktuellen Stand der Technik vergleichbar. Die Phasen laufen nach bisherigen Erkenntnissen überwiegend sequenziell ab, was für eine automatisierte Erkennung vorteilhaft ist. Lediglich Phase 7 („Blutstillung & Spülung“) tritt u. U. auch während oder

zwischen anderen Phasen auf. Die beiden Abbildungen 4.9 und 4.10 zeigen die Verteilung und die Übergangswahrscheinlichkeiten der Phasen im Datenset des HeiChole Benchmarks, was die vorherige Aussage stützt. Gleichzeitig ist erkennbar, dass eine gewisse Unsicherheit im Ablauf der Phasen vorherrscht, sodass nicht immer von der gleichen Reihenfolge ausgegangen werden kann.

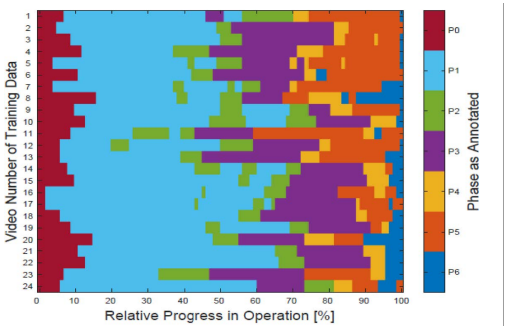


Abbildung 4.9: Phasenverteilung im Trainingsdatensatz. Die Dauer des Vorgangs ist auf 100% normiert und die Phasenannotationen werden in Schritten von 1% dargestellt. Folgende Phasen werden angezeigt: Vorbereitung (P0), Dissektion Calot Dreieck (P1), Clippen und Schneiden (P2), Dissektion Gallenblase (P3), Verpackung Gallenblase (P4), Blutstillung und Spülung (P5) und Bergung Gallenblase (P6) [159].

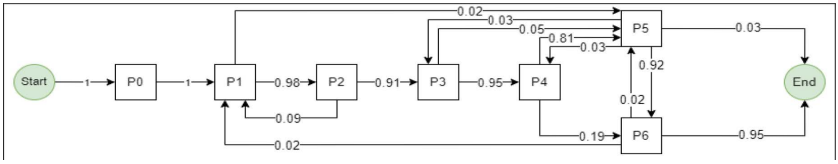


Abbildung 4.10: Grafische Darstellung der Phasen und ihrer möglichen Übergänge. Die Wahrscheinlichkeiten der Phasenübergänge wurden auf der Grundlage des gesamten Datensatzes (Trainings- und Testdatensatz zusammen) berechnet [159].

Die Phaseneinteilung soll letztendlich auch Rückschlüsse auf den Operationsverlauf zulassen, weshalb eine zeitliche Abschätzung der Einzelschritte und somit der jeweiligen Phase mit in die Untergliederung einfließt. Die Zeitangaben erfolgen nach verschiedenen Literaturangaben ([74], [93], [94], [120] & [151]).

Tabelle 4.1: Operationsschritte der laparoskopischen Cholezystektomie.

Periop. Phase	Operationsschritte	Zuordnung im eigenen Phasenmodell	Dauer (min)
präoperativ	1. Patient in OP-Saal bringen 2. Patient lagern 3. Schnittstellen desinfizieren 4. Patient steril abdecken	1. präoperativ	39 ± 21
intraoperativ	5. Schnitt am Nabelrand 6. Pneumoperitoneum anlegen 7. Kameratrokar einbringen 8. Inspektion der Bauchhöhle 9. Arbeitstrokare und Instrumente einbringen	2. Vorbereitung	$21{,}5 \pm 13{,}5$
	10. Gallenblase fassen & hochschieben 11. Ductus cysticus & Arteria cystica darstellen 12. Gallenblase freipräparieren	3. Dissektion Calot Dreieck	16 ± 9
	13. Clippen & Schneiden des Ductus cysticus 14. Clippen & Schneiden der Arteria cystica	4. Clippen & Schneiden	$3 \pm 2{,}5$
	15. Gallenblase fassen 16. Gallenblase aus Gallenblasenbett ausschälen	5. Dissektion Gallenblase	14 ± 9
	17. Gallenblase in Beutel packen	6. Verpackung Gallenblase	$1{,}5 \pm 1$
	18. Blutstillung & Spülung	7. Blutstillung & Spülung	$3 \pm 2{,}5$
	19. Gallenblase bergen	8. Bergung Gallenblase	6 ± 3
	20. Untersuchung auf Blutungen 21. Oberbauch reinigen & Spülen 22. Lagekontrolle der Clips 23. Trokare entfernen 24. Ablassen des Pneumoperitoneums 25. Naht	9. Abschluss	16 ± 4
postoperativ	26. Patient entlagern 27. Patient extubieren 28. Patient ausleiten	10. postoperativ	$30{,}5 \pm 4{,}5$

Wie in Abschnitt 2.1.4 bereits erläutert, hängen die Zeiten von diversen Faktoren ab, was sich auch in den teilweise großen Streubreiten der angegebenen Phasendauern widerspiegelt. So schwankt, laut Tabelle 4.1, die Dissektion der Gallenblase zwischen 5 und 23 Minuten. Aus diesem Grund sind die hier dargestellten Angaben eher als grobe Richtwerte und weniger als konkrete Vorgaben zu sehen. Trotzdem ermöglicht diese Aufstellung eine Übersicht der zeitlichen Verteilung der einzelnen Phasen im Gesamtprozess und dadurch auch zumindest eine relative Abschätzung des zeitlichen Verlaufs laufender OPs.

4.2.2 Sensorische Erfassung der identifizierten OP-Prozesse

Nachdem die Prozessabläufe während der laparoskopischen Cholezystektomie erläutert wurden, untersucht der folgende Abschnitt Möglichkeiten zur sensorischen Erfassung der einzelnen OP-Schritte und -phasen. Da einerseits die größten Unsicherheiten, zeitlichen Varianzen und entsprechend schlechte Planbarkeit bei den intraoperativen Prozessen vorliegen und andererseits die kontrollierte Umgebung innerhalb des OP-Saals eine systematische Auswertung erheblich vereinfacht, liegt der Fokus der nachfolgenden Arbeiten auf den Abläufen aus Tabelle 4.1. Dabei wird, unter Berücksichtigung der in Abschnitt 4.1 bereits definierten Anforderungen, insbesondere analysiert, was die für den spezifischen Abschnitt relevanten Merkmale sind, mit welchen Methoden diese gemessen werden können und welche Sensorik dafür geeignet ist. So wird bspw. körpernahe Sensorik am OP-Team oder am Patienten nicht näher betrachtet, da dies gegen die Anforderungen [*AF-10*] und [*AF-12*] verstößt. Weiterhin wird auch Sensorik ausgeschlossen, die Veränderungen an den bestehenden OP-Geräten und Instrumenten erfordert ([*AF-09*], [*AF-10*] & ggf. [*AF-13*]). Die Granularität der Analyse folgt den Phasen des intraoperativen Prozesses ergänzt durch die präoperative Phase, den OP-Abschluss und die postoperative Phase (vgl. Spalte „Zuordnung im eigenen Phasenmodell“ in Tabelle 4.1).

Das Ergebnis der Analyse ist in Tabelle 4.2 dargestellt. Es fällt auf, dass beginnend in Phase 2 und endend in Phase 9, also nahezu der gesamte Anteil des intraoperativen Abschnitts, die Operation laparoskopisch, also im Körperinneren des Patienten, stattfindet. Während dieser Zeit findet außerhalb des Patienten nur sehr wenig Aktivität statt. Lediglich bei Instrumentenwechseln können Veränderungen, insb. am Instrumententisch, festgestellt werden. Zur Erkennung dieser Phasen wird also Sensorik mit entsprechenden Methoden benötigt, die Analysen im Körperinneren und am Instrumententisch ermöglichen. Ohne Änderungen an den eingesetzten Instrumenten, die unter Beachtung regulatorischer Anforderungen ohne den jeweiligen Hersteller nicht möglich sind und dadurch

Anforderung [*AF-13*] widerspricht, sind die Optionen zur Erfassung der Aktivität im Körperinneren beschränkt. Eine Auswertung der Bild- bzw. Videodaten aus der Endoskopkamera bildet den naheliegendsten Lösungsansatz. Ähnlich verhält es sich bei der Analyse am Instrumententisch. Direkte Messmethoden zur Unterscheidung einzelner Instrumente und zur Erfassung ihrer Nutzung, bspw. über RFID, erfordern Änderungen der Instrumente selbst und/oder des Instrumententischs, was aus zuvor genannten Gründen aus der Entscheidungsfindung ausgeschlossen wird. Als Alternative kann auch hier ein Kamerasystem zur Analyse des Instrumententischs eingesetzt werden. Durch ein desinfizierbares Gehäuse und geschickte Installation können die genannten Probleme vermieden werden. Darüber hinaus ist der Informationsgehalt in Kamerabildern enorm hoch, sodass dadurch verschiedene Parameter, wie bspw. die Unterscheidung von Instrumenten, der jeweilige Nutzungszeitpunkt, Veränderungen am Instrument (z. B. Verschmutzung) und weitere gewonnen werden können. Auch in der präoperativen Phase kommen teilweise Instrumente und Materialien vom Instrumententisch zum Einsatz (z. B. Tupferklemme oder sterile Tücher), damit das Kamerasystem zur Beobachtung des Instrumententischs sinnvoll verwendet werden könnte. Aus Tabelle 4.2 wird außerdem ersichtlich, dass bestimmte Materialien spezifisch für einzelne Phasen sind (bspw. Tupferklemme & sterile Tücher in Phase 1, Skalpell, Veress-Nadel & Trokare in Phase 2, Clipper & Schere in Phase 4, usw.). Dies bestärkt ein Instrumententracking-System als sinnvolle Komponente zur Phasenerkennung. Es wird allerdings auch deutlich, dass der Informationsgehalt der Nutzung einzelner Instrumente alleine wahrscheinlich zu gering für eine robuste Erkennung ist. So ist der Greifer in jeder intraoperativen Phase sichtbar und birgt somit keinen Mehrwert. Auch der Haken kommt in beiden Dissektionsphasen zum Einsatz, was die Differenzierung durch reines Instrumententracking verhindert. Außerdem sind die genutzten Instrumente bei den Phasen 6 und 8 („Verpackung Gallenblase“ und „Bergung Gallenblase“) identisch (Greifer + Beutel). Bei der Bergung allerdings ist der Beutel gefüllt, was ggf. ebenfalls erkannt werden könnte. Auch die Analysen in den Abbildungen 4.8a bis 4.8d zur Instrumentennutzung in realen Szenarien zeigt einige Unsicherheiten und entsprechend nicht eindeutige Zuordnungen zwischen den

eingesetzten Instrumenten und den OP-Phasen, was darauf hindeutet, dass die Instrumentennutzung alleine nicht für die Erkennung der OP-Phasen ausreicht.

In den Phasen 1, 2, 9 und 10 finden die meisten Arbeiten außerhalb des Patientenkörpers statt, sodass sich hier weitere Analysemöglichkeiten ergeben. So lässt in den genannten Phasen die Anwesenheit und Positionierung bestimmter Personen oder Personengruppen Rückschlüsse auf den Operationsverlauf zu. In Phase 1 ist z. B. das Erscheinen und Positionieren des Patienten entscheidend. In den Phasen 2 und 9 hingegen betreten bzw. verlassen die Chirurgen und deren Assistenten den Raum. Die Erfassung von Personen sowie die Erkennung ihrer Positionierung und (Lauf-)Wege kann ebenfalls mit Hilfe von Bildanalysemethoden erfolgen, sodass auch diese Informationen aus Kamerasystemen gewonnen werden können. In diesem Fall müsste der gesamte OP-Saal erfasst werden, um ein vollständiges Bild zu erhalten.

Zusätzliche Informationen können durch die Vernetzung des zu entwickelnden Systems mit verschiedenen Geräten im OP-Saal generiert werden. Beispiele hierfür sind der Gasfluss im Insufflator, Einstellungen am OP-Tisch und der Lichtquelle oder die Nutzung des Hochfrequenz (HF)-Geräts, der Spülung und der Absaugung. In einem hochintegrierten OP-Saal stehen diese Informationen, entsprechende Schnittstellen zur Integration neuer Systeme vorausgesetzt, implizit zur Verfügung.

Tabelle 4.2: Analyse der Operationsschritte der laparoskopischen Cholezystektomie.

Phase	signifikante Arbeitsschritte	relevante Materialien	messbare Parameter	Analysemethodik	Sensorik
1. Präoperativ	Patient kommt in Saal, Desinfektion, sterile Abdeckung	Tupferklemme, Tupfer, Abdecktücher (farbig, steril)	Anwesenheit Patient, Instrumentennutzung, desinf. Körperstellen, Tücher	Personen-, Instrumenten- & Materialerkennung/-Tracking	Näherungssensorik, Raum-/Tisch-kamera
2. Vorbereitung	Neigung OP-Tisch, Positionierung Personal, intraabd. Druckaufbau, Schnitt, Abdunklung, Sicht Bauchraum	Skalpell, Veress-Nadel/-Kanüle, Kamera & Arbeitstrokare, Greifer	Personalbewegung, Instrumentennutzung, Gasfluss Insufflator, Helligkeitsänderung Raum, Einstellung OP-Tisch	Personen- & Instrumententracking, Vernetzung OP-Geräte (Insufflator, OP-Tisch)	Raum-/Tisch-/Endoskopkamera, Helligkeitssensor, Neigungssensor
3. Dissektion Calot Dreieck	Laparoskopie	Greifer, Haken	Instrumentennutzung	Instrumententracking	Tisch-/Endoskop-kamera
4. Clippen & Schneiden	Laparoskopie	Greifer, Clipper/ Stapler, Schere	Instrumentennutzung	Instrumententracking	Tisch-/Endoskop-kamera
5. Dissektion Gallenblase	Laparoskopie	Greifer, Haken	Instrumentennutzung	Instrumententracking	Tisch-/Endoskop-kamera
6. Verpackung Gallenblase	Laparoskopie	Greifer, Beutel	Instrumentennutzung	Instrumententracking	Tisch-/Endoskop-kamera
7. Blutstillung & Spülung	Laparoskopie	Greifer, Bipolar, Saug-Spül-Rohr	Instrumentennutzung	Instrumententracking, Vernetzung OP-Geräte (HF-Gerät, Spülung/Absaugung)	Tisch-/Endoskop-kamera
8. Bergung Gallenblase	Laparoskopie, Beutel Bauchraum → Instrumententisch	Greifer, Beutel (gefüllt, verschlossen)	Instrumentennutzung	Instrumententracking	Tisch-/Endoskop-kamera
9. Abschluss	Ende Nutzung endosk. Geräte, Chirurg & Assistenten verlassen Raum	Pinzette, Nadelhalter, Nadel, Fäden	Instrumentennutzung, Abwesenheit Personal	Personen- & Instrumententracking	Raum-/Tisch-/Endoskopkamera
10. Postoperativ	Tischneigung in Ausgangslage, Extubation, Patient aus OP-Saal	-	Einstellung OP-Tisch, Abwesenheit Personal/Patient	Vernetzung OP-Geräte (OP-Tisch), Personentracking	Raumkamera, Neigungssensor

Sowohl ein Neigungssensor am OP-Tisch, ein Näherungssensor zur Erfassung bestimmter Positionierungen des Personals als auch ein Helligkeitssensor bieten, im Vergleich zu Kameratechnik in Ergänzung mit der Vernetzung der Geräte, lediglich redundante Informationen. Gleichzeitig erfordern sie aber auch erhöhten Installations- und Pflegeaufwand. Bei der weiteren Konzeptionierung des Gesamtsystems muss entsprechend zwischen Kosten und Nutzen dieser zusätzlichen Sensorik abgewogen werden. An dieser Stelle der Arbeit ist davon auszugehen, dass mithilfe der automatisierten, videobasierten Objekt- und Personenerkennung über Videokameras alle notwendigen Prozesse für den Rückschluss auf die OP-Phasen erkennbar sind. Dennoch bleibt der Einsatz weiterer Sensoren sowie die Kombination von Sensorik mit dem Kamerasystem als offene Möglichkeit bestehen und soll die Handlungserkennung an den notwendigen Stellen vereinfachen.

Zusammenfassend gilt für die durchgeführte Analyse, dass die Grundfunktionalität des Erkennungssystems alleinig durch den Einsatz von Kameras als Sensorik umgesetzt werden kann. Weitere Erkenntnisse werden in den spezifischen Anforderungen im folgenden Abschnitt erläutert.

4.2.3 Spezifische Anforderungen zur sensorischen Erkennung von OP-Phasen

Aus den tiefergehenden Analysen der laparoskopischen Cholezytektomie bzgl. der Abläufe und Prozesse sowie deren Erkennbarkeit ergeben sich zusätzlich zu den allgemeinen Anforderungen aus Abschnitt 4.1 folgende spezifischen Anforderungen:

[*S-AF-01*] Das System soll Zugriff auf andere Netzwerkgeräte haben, um Teil des integrierten OPs zu werden.

[*S-AF-02*] Das System muss Informationen aus dem Körperinneren des Patienten zur Analyse des OP-Verlaufs auswerten können.

[*S-AF-03*] Ein Kamerasystem muss so angebracht sein, dass der gesamte OP-Saal erfasst wird. Weiterhin muss die Bildqualität bzgl. Auflösung, Kontrastverhältnis und Lichtempfindlichkeit ausreichend sein, um Personen unterscheiden und deren Bewegungen erkennen zu können.

[*S-AF-04*] Ein Kamerasystem muss so angebracht sein, dass der komplette Instrumententisch erfasst wird. Weiterhin muss die Bildqualität bzgl. Auflösung, Kontrastverhältnis und Lichtempfindlichkeit ausreichend sein, um eine Unterscheidung der auf dem Tisch befindlichen Instrumente zu ermöglichen.

[*S-AF-05*] Das System soll bei Bedarf um weitere Sensorik erweiterbar sein.

4.3 Konzeption des Gesamtsystems

Basierend auf den bisher gewonnenen Erkenntnissen wird im folgenden Abschnitt das Gesamtkonzept des Erkennungssystems erarbeitet. Die Grundidee hierfür ist, dass für die jeweiligen zu erkennenden Operations-Phasen typische Merkmale erfasst werden und dadurch ein kontinuierliches Bild über laufende und abgeschlossene Phasen entsteht. Die Merkmale können, wie in Abschnitt 4.2.2 diskutiert, bspw. die Anwesenheit bestimmter Personen, der Nutzungszeitpunkt bestimmter Instrumente und Materialien oder spezifische Einstellungen an den OP-Geräten sein. Gemäß den Ergebnissen aus Abschnitt 4.2.2 liegt der Schwerpunkt der Sensorik und Methodik des Gesamtsystems für diese Arbeit dabei auf videobasierter Aktivitätserkennung. Dabei werden das Tracking der Instrumente, sowohl am Instrumententisch, als auch im Bauchraum des Patienten und das Tracking von Personen als besonders vielversprechende Informationsquellen erachtet.

Das Diagramm in Abbildung 4.11 skizziert den Gesamtablauf des Erkennungssystems. Die Phasen sind hierbei als Blöcke dargestellt, die verschiedene Aktivitäten und Zustände inkludieren. Diese wiederum sind als längliche Ovale

dargestellt. Die unterschiedlichen Farben verdeutlichen, durch welches Teilsystem (Deckenkamera, Instrumententischkamera, Endoskopkamera oder Gerätevernetzung) diese Aktivität im vorliegenden Konzept erkannt werden soll. Einige Zustände können prinzipiell von verschiedenen Teilsystemen bestimmt werden. Bspw. erkennt sowohl die Deckenkamera als auch die Vernetzung mit der Beleuchtung den abgedunkelten Raum. Für die Darstellung in Abbildung 4.11 wurde in diesen Fällen jeweils das vermeintlich geeignetere System ausgewählt, wobei darauf geachtet wurde, möglichst viele Punkte über die verschiedenen Kamerasysteme abzudecken. In der Umsetzung können die redundanten Informationen dann zur Verbesserung der Erkennungsqualität beitragen. Die Ablaufdiagramme der einzelnen Phasen sind für eine bessere Lesbarkeit in Anhang A.1.2 nochmal separat aufgeführt.

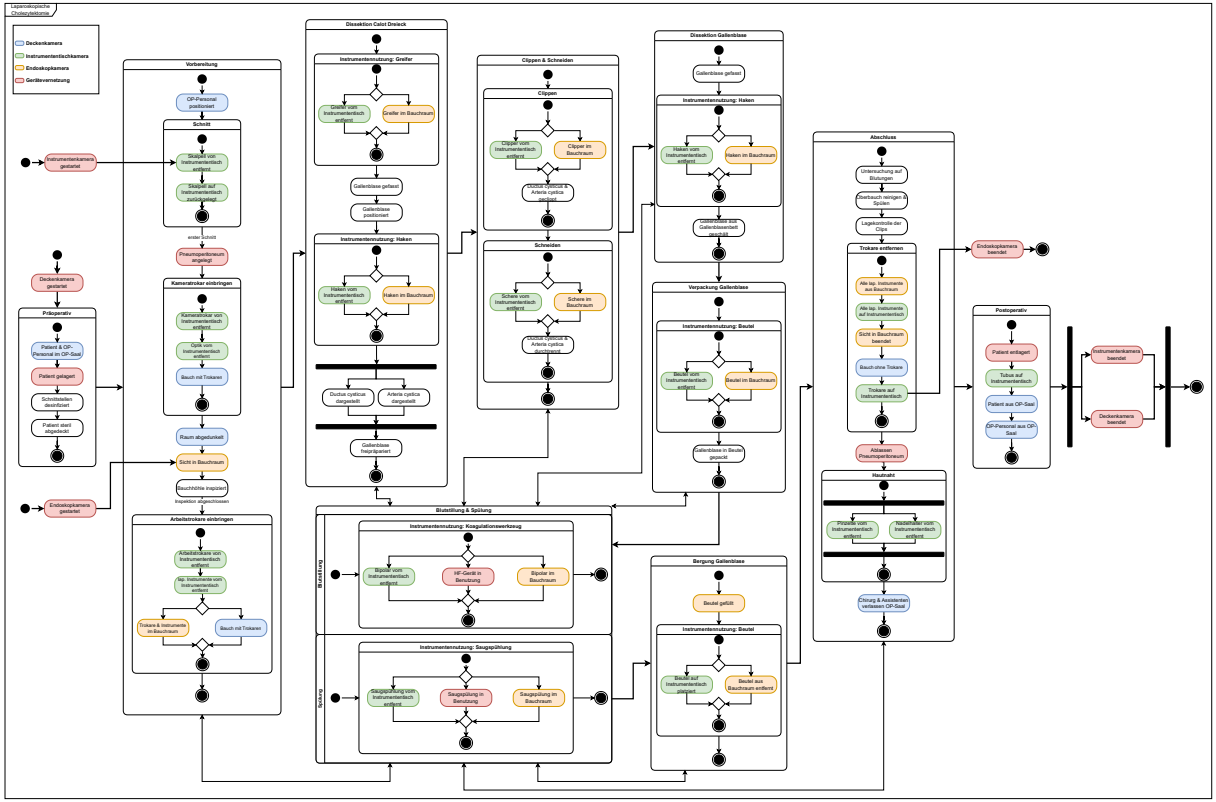


Abbildung 4.11: Ablaufdiagramm des Gesamtsystems. Zur Betrachtung wird die digitale Version dieses Dokuments empfohlen.

Zunächst wird nach den bisherigen Erkenntnissen deutlich, dass die einzelnen Phasen überwiegend sequenziell ablaufen. Lediglich Phase 7 „Blutstillung & Spülung“ kann auch zwischen den übrigen Phasen auftreten. Im Standardablauf findet sie allerdings zwischen den Phasen „Verpackung Gallenblase“ und „Bergung Gallenblase“ statt. Der sequenzielle Ablauf kann, basierend auf diesen Annahmen, für die automatische Phasenerkennung ausgenutzt werden. Weiterhin wird deutlich, dass die unterschiedlichen Teilsysteme zu unterschiedlichen Zeitpunkten starten und enden. So kommen bspw. die Endoskop- und die Instrumententischkamera erst in der Vorbereitungsphase zum Einsatz, wohingegen die Deckenkamera bereits in der präoperativen Phase und die Gerätevernetzung sogar noch davor Verwendung finden. Letztere ist somit geeignet, um den Start und auch das Ende der Erkennung festzulegen.

Wie bereits zuvor beschrieben, ist in den intraoperativen Phasen hauptsächlich die Erkennung der Instrumentennutzung als Informationsquelle über den OP-Verlauf verfügbar. Hierfür können die Endoskop- und die Instrumententischkamera simultan eingesetzt werden. Dies ermöglicht durch die Redundanz einerseits eine erhöhte Erkennungskonfidenz und durch die unterschiedlichen Blickwinkel und Fokusse auf die Instrumente andererseits auch eine bessere Diskriminierungswahrscheinlichkeit insb. von ähnlichen Instrumenten. Die Deckenkamera dient vor allem der Erkennung von anwesenden Personen(gruppen) (z. B. Patient oder Chirurg) und deren Verlassen des OP-Saals sowie deren Aktivitätsniveau. Letzteres wird durch die „Menge der Bewegung“ erfasst (Details in Kapitel 5). Über die Gerätevernetzung soll der Zustand bestimmter Geräte im OP-Saal erfasst und somit auf deren Nutzung (z. B. HF-Gerät bei Koagulation) oder Einstellungsänderungen (z. B. Helligkeitsveränderung der Lichtquelle oder Stellung des OP-Tischs) geschlossen werden. Verglichen mit den Analysen in den Tabellen 4.1 und 4.2 sind im dargestellten Ablauf nicht alle Einzelschritte bzw. deren Merkmale erfasst. Dies ist zum einen in der nur schwer umsetzbaren technischen Erfassbarkeit mancher Merkmale und zum anderen in der damit verbundenen Abwägung von Aufwand und Nutzen im Gesamtsystem begründet. Abbildung 4.11 verdeutlicht jedoch, dass für jede Phase signifikante Merkmale zur eindeutigen Identifikation definiert wurden.

Aus diesem Systemkonzept ergibt sich zusammen mit den Ergebnissen der Anforderungsanalyse die in Abbildung 4.12 dargestellte Systemskizze. Diese gliedert sich in drei Blöcke: Kamerasystem, KIS und OP-Geräte.

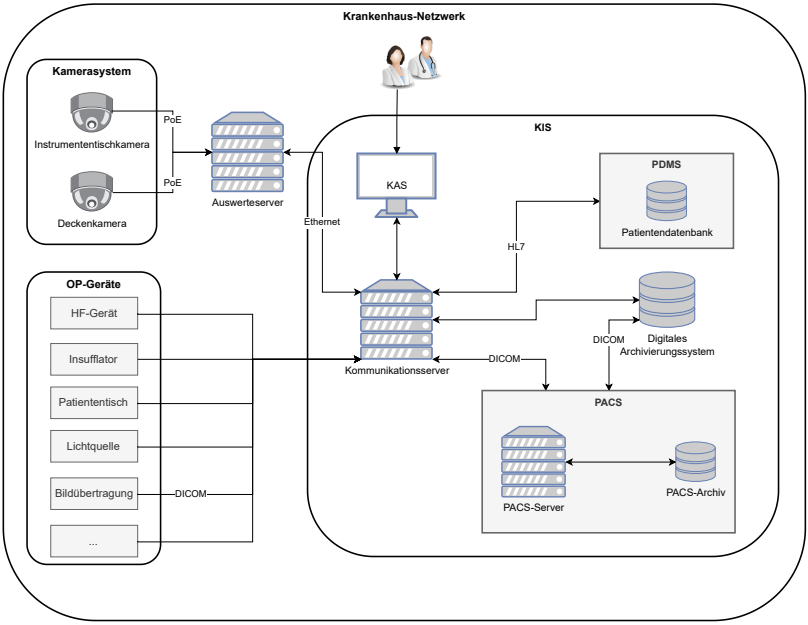


Abbildung 4.12: Systemskizze des Gesamtsystems

Der Fokus dieser Arbeit liegt auf den Kamerasystemen und den OP-Geräten. Ein Grundverständnis über das KIS ist dennoch wichtig, um die Einbettung und Abgrenzung der in dieser Arbeit vorgestellten Komponenten innerhalb der Krankenhaus-IT zu durchblicken. Für die Entwicklung des Gesamtsystems sind für die Blöcke KIS und OP-Geräte die jeweiligen Schnittstellen zu bedienen. Diese hängen von den verwendeten Systemen des jeweiligen Krankenhauses ab und können nicht allgemeingültig benannt werden. Es ist aber davon auszugehen, dass auf übliche Standards wie HL7 und DICOM zurückgegriffen werden kann.

Das Kamerasystem besteht nach diesem Konzept aus zwei unabhängigen Netzwerkkameras. Die Auswahl von sog. Internet Protocol (IP)-Kameras ermöglicht sowohl die direkte Einbindung ins Krankenhausnetzwerk und somit eine einfache Integration der Daten in bestehende Systeme als auch einen geringen Installationsaufwand im OP-Saal durch die Versorgung über Power over Ethernet (PoE) (erfüllt Anforderung [*AF-09*]). Eine der Kameras übernimmt dabei die Funktion der Beobachtung des Instrumententischs (Anforderung [*S-AF-04*]), die zweite Kamera dient der Raumbeobachtung (Anforderung [*S-AF-03*]). Technisch könnten auch beide Aufgaben von einer Kamera übernommen werden. Eine Aufteilung in separate Systeme erlaubt allerdings höhere Flexibilität und Robustheit des Gesamtsystems, bspw. ggü. Verdeckung, sowie eine Entlastung der Auswerteeinheit, da zusätzliche Operationen wie Cropping der ROI zur Fokussierung auf den Instrumententisch entfallen. Für die Entwicklungen in dieser Arbeit wurde für die Deckenkameras jeweils das Modell Canon VB-H45 ausgewählt. Diese kommt auch im Richard Wolf eigenen OP-Integrationssystem core nova[1] zum Einsatz, was sicherstellt, dass sie für die klinische Umgebung geeignet ist. Außerdem bietet sie eine Motorsteuerung, um stets die gewünschte Ausrichtung zu erreichen sowie eine Lichtempfindlichkeit von 0,05 Lux bei F1,6/50 IRE, um auch bei abgedunkeltem Raum auswertbare Bilder aufzeichnen zu können und dadurch Anforderung [*AF-11*] sowie den Risiken aus Abschnitt 4.4.1 gerecht zu werden. Weitere Eigenschaften dieses Modells finden sich im Anhang A.2. Die Auswertung aller anfallenden Informationen findet auf einem zentralen Server statt. Dies ermöglicht eine gute Skalierbarkeit auch bei mehreren OP-Sälen.

Um die einzelnen Teilsysteme zu einem Gesamtergebnis im Hinblick auf die Erkennung von OP-Phasen für die Optimierung der chirurgischen Logistik auswerten zu können, ist ein multimodales Multitask-Modell erforderlich. Abbildung 4.13 skizziert eine dafür geeignete Gesamtarchitektur, welche für die vorliegende Arbeit konzipiert wurde. Das dargestellte Konzept sieht dabei zunächst ein separates Modell für jedes Teilsystem vor. Die Ergebnisse dieser Modelle

[1] `https://www.riwolink.com/en/solutions/integration`, zuletzt geprüft: 10.02.2024

fließen anschließend als Eingabedaten in ein übergeordnetes Auswertemodul, welches das finale Analyseergebnis berechnet. Vorteil dieses Ansatzes im Vergleich zu anderen Multitask-Ansätzen, die bspw. ein früheres Parameter-Sharing ermöglichen oder End-to-End trainiert werden können, ist die höhere Flexibilität. So können die einzelnen Modelle der Teilsysteme unabhängig voneinander implementiert werden, was u. A. die Beschaffung der jeweiligen Trainingsdaten vereinfacht. Weiterhin können die Zwischenergebnisse auch anderweitig verwendet werden, was auch Auswertungen, die keinen direkten Bezug zur OP-Logistik aufweisen, erlaubt. Beispiele hierfür sind Maßnahmen der Patientensicherheit, wie das Zählen von Materialien in und aus dem Körper durch die Instrumentenerkennung, oder Ergonomieoptimierungen auf Basis von Haltungsanalysen des OP-Personals mittels Skelettanalysen. Zudem ermöglicht dieser Ansatz eine Erweiterung des Gesamtmodells, ohne dass die bestehenden Teilsysteme neu trainiert werden müssen (Anforderung [*S-AF-05*]).

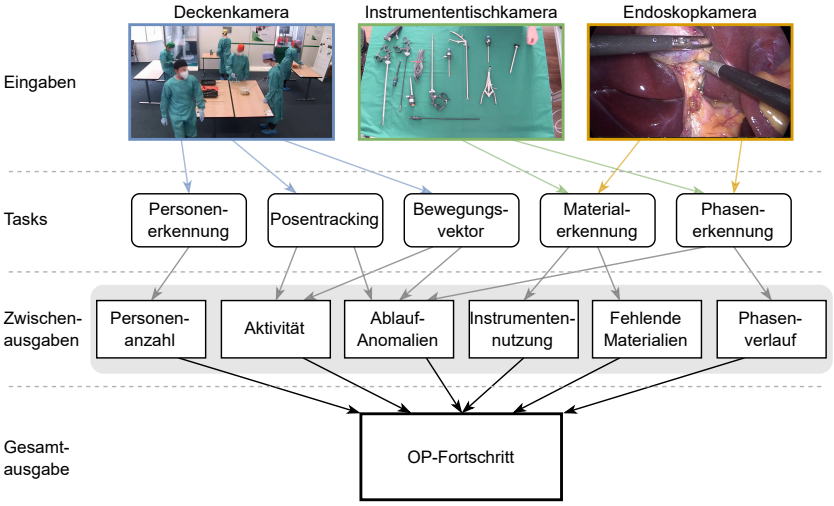


Abbildung 4.13: Gesamtmodell zur Auswertung von Operationen für die Optimierung der chirurgischen Logistik (Bildquelle Endoskopkamera: [159])

4.4 Risikobetrachtung

Im klinischen Umfeld ist eine Risikoanalyse neuartiger Entwicklungen unumgänglich. Im Rahmen dieser Arbeit wurden unterschiedliche Risiken identifiziert. Diese können grob in vier Kategorien (Kamerasystem, Schnittstellen, Algorithmik, Gesamtsystem) eingeteilt werden, wie in den nachfolgenden Abschnitten detailliert dargestellt.

4.4.1 Kamerasystem

Kamerabasierte Systeme bergen immer Risiken hinsichtlich verschiedener Umgebungseinflüsse. Ein wichtiger Punkt dabei ist die Beleuchtung bzw. die Lichtempfindlichkeit des Kamerasensors. Fällt diese zu gering aus, werden Bilder nicht ausreichend belichtet und sind zu dunkel, um etwas zu erkennen. Üblicherweise herrscht in OP-Umgebungen eine gleichbleibende und gleichmäßig verteilte Beleuchtung. Gerade während laparoskopischen Eingriffen muss der Raum allerdings aufgrund der eingesetzten Monitortechnik stark abgedunkelt werden. Hierbei konnte in einer Stichprobe ein Wert von lediglich 4 Lux gemessen werden. Dies muss bei der Auswahl der Kameras Beachtung finden. Bei Endoskopkameras besteht diesbezüglich kein Risiko, da diese speziell für den Anwendungsfall konzipiert sind und über eine zusätzliche Lichtquelle verfügen.

Ein weiteres Risiko beim Einsatz von Kamerasystemen besteht durch Verdeckung (bspw. durch Personal oder Instrumente) oder insbesondere bei Endoskopkameras durch Verschmutzung der Linsentechnik (bspw. durch Blut). Dies lässt sich nicht komplett vermeiden, eine geschickte Platzierung sowie redundante Aufnahmen aus unterschiedlichen Blickwinkeln können das Risiko aber minimieren. Bei Verschmutzung muss die Linse gesäubert werden. In gewissem Rahmen können die Erkennungsalgorithmen nicht vollständige Verdeckungen aber auch auflösen.

4.4.2 Schnittstellen zu anderen Systemen

Da das geplante System auf bestehende Systeme (KIS, OP-Instrumente) zugreifen soll, benötigt es entsprechende Schnittstellen. Sollten diese nicht verfügbar sein, kann dieser Teil nicht umgesetzt werden. Auf Seiten der OP-Instrumente steht mit der Richard Wolf GmbH ein Hersteller als Ansprechpartner zur Verfügung, sodass lediglich ein äußerst niedriges Risiko bzgl. fehlender Schnittstellen besteht. Es bleibt allerdings ein Restrisiko, wenn ein potentielles Partnerkrankenhaus andere Hersteller nutzt. Dann wäre eine Zusammenarbeit mit diesen Herstellern ebenfalls notwendig.

Beim Zugriff auf das KIS besteht ebenfalls eine starke Abhängigkeit zum Partnerkrankenhaus, da verschiedene KIS-Hersteller auf unterschiedliche Schnittstellen setzen. Diese sind meist bekannt und gut dokumentiert (z. B. HL7). Allerdings ist deren Anbindung häufig mit nicht unerheblichen Kosten und Aufwand verbunden. Im Rahmen dieser Arbeit wird der Einbezug des KIS nicht weiter betrachtet. Gründe dafür sind einerseits der Fokus auf die Analyse von Videodaten als Erkenntnis aus den vorangegangenen Abschnitten und andererseits die genannten Risikofaktoren. Für zukünftige Weiterentwicklungen und einen etwaigen Produktiveinsatz wäre das jedoch ein zu beachtendes Thema.

4.4.3 Algorithmen & Hardware

Die Anforderungen an das geplante System sind sehr hoch. So muss die Berechnungsgeschwindigkeit hoch genug sein, um eine möglichst geringe Verzögerung zu erreichen, da andernfalls der Mehrwert des Systems entfällt. Hierfür wird besonders leistungsfähige Hardware benötigt. Insbesondere spezialisierte Grafikbeschleuniger können die Berechnungen beschleunigen. Dies muss bei der Systemspezifikation beachtet werden. Auf algorithmischer Seite ist auch ein mehrstufiger Prozess denkbar, bei dem zeitkritischere Komponenten schneller berechnet und weniger kritische Teile verzögert abgearbeitet werden.

4.4.4 Daten für das Training von maschinellen Lernmethoden

Für das Training von Modellen mittels maschinellem Lernen, insb. im Bereich des überwachten Lernens, werden üblicherweise große Mengen an annotierten Daten benötigt. Diese müssen sowohl repräsentativ für die zu erfüllende Aufgabe als auch in ausreichender Qualität und Diversität vorliegen. Für die Annotationen ist, neben dem technischen Wissen zur Erfassung der Daten, das jeweilige Domänenwissen notwendig. Alle diese Komponenten erzeugen zeitlichen wie auch technischen und somit monetären Aufwand.

Darüber hinaus sind die Daten vor allem im Gesundheitswesen oftmals sensibel und bedürfen besonderer Aufmerksamkeit bzgl. Datenschutz und Datensicherheit. Dies erschwert zusätzlich die Beschaffung und Nutzung der Daten für die Modellentwicklung. Aus diesem Grund sind, anders als in anderen Domänen auch nur wenige öffentliche Datensets für medizinische Anwendungen verfügbar. Dies führt dazu, dass meist eigene Daten erzeugt werden müssen. Dabei sind entsprechende Maßnahmen zum Datenschutz und Datensicherheit sowie rechtliche Aspekte umzusetzen. Beispiele für die Umsetzung der genannten Aspekte sind die Unkenntlichmachung personenbezogener Daten, eingeschränkter Zugriff nur für berechtigte Personen oder Datenschutzvereinbarungen mit den Datengebenden.

Des Weiteren entstehen zusätzliche Abhängigkeiten, da für die Installation zusätzlicher Sensorik und deren Anbindung an die Krankenhaus-IT sowohl die Zustimmung der Einrichtung gegeben, als auch das Personal zur Einbindung verfügbar sein muss. Zugleich ist die Qualität und Quantität der Daten auch abhängig von der Häufigkeit des aufzuzeichnenden Ereignisses, im vorliegenden Fall also konkret die Häufigkeit laparoskopischer Cholezystektomien in der Einrichtung.

Insgesamt besteht also ein großes Risiko darin, überhaupt geeignete Daten für die Umsetzung des Vorhabens zu beschaffen, während erhebliche Verzögerungen

in der Planung aufgrund vieler unterschiedlicher Abhängigkeiten, die nicht komplett beeinflussbar sind, einen weiteren Risikofaktor darstellen.

4.4.5 Gesamtsystem & Infrastruktur

Bzgl. des Gesamtsystems bestehen einerseits infrastrukturelle und andererseits personelle Risiken. Bei der Infrastruktur stellt sich die Frage, ob die Berechnungen lokal innerhalb des OP-Saals stattfinden oder die Daten zunächst in ein (möglicherweise selbst gehostetes) Rechenzentrum gestreamt werden. Letzteres bietet den Vorteil, dass wesentlich mehr Rechenleistung zur Verfügung gestellt werden kann, bzw. diese auch besser auf sich ändernde Anforderungen skalierbar ist. Auf der anderen Seite muss die Infrastruktur zum schnellen Datenaustausch bereitgestellt werden, wodurch allerdings neue Risiken bspw. bzgl. Ausfall- oder Datensicherheit entstehen. Eine lokale Berechnung hingegen erfordert eine Bereitstellung entsprechender Hardwarekomponenten vor Ort. Eine Zwischenlösung stellt die Platzierung der Hardware in einem Nebenraum dar, zu welcher eine stabile Verbindung aufgebaut werden kann, die Komponenten sich aber außerhalb des OP-Bereichs befinden.

Ein letztes erwähnenswertes Risiko ist mangelnde Nutzerakzeptanz. Mitarbeitende stehen Systemen, welche Bilddaten aufnehmen und auswerten, häufig skeptisch gegenüber. Grund hierfür sind Bedenken bzgl. Datenschutz und Datensicherheit. Dem kann, nach persönlicher Erfahrung des Autors, mit sorgfältiger Aufklärung zu Wirkungsweise und Nutzen des Systems entgegengewirkt werden.

4.5 Ablauf zur Umsetzung des Konzepts

Die in diesem Kapitel erläuterten Anforderungen und Konzepte schaffen die Grundlage für die nachfolgende Entwicklung des Systems zur OP-Phasenerkennung. Dabei wurden insbesondere drei verschiedene Kamerasysteme als Basis für die geplante Analyse von OP-Abläufen definiert, was den algorithmischen

Schwerpunkt der Arbeit auf Methoden aus den Bereichen des maschinellen Sehens und der videobasierten Aktivitätserkennung legt. Zusätzlich wurden Risiken verschiedener Komponenten betrachtet, die ebenfalls mit in den Entwicklungsprozess einfließen. Die Konzeption und Umsetzung der dargestellten Teilsysteme wird in den folgenden Kapiteln detailliert erläutert.

Kapitel 5 beschreibt die Entwicklung des Teilsystems zur Beobachtung des gesamten Raumes mittels Deckenkamera unter Verwendung eines skelettbasierten Ansatzes. Hierbei werden außerdem die besondere Herausforderung der Re-Identifikation von Personen im medizinischen Kontext und mögliche Lösungsansätze diskutiert (5.3). In Kapitel 6 wird das Teilsystem zum Tracking der Materialien am Instrumententisch durch Objekterkennungsmethoden näher erläutert, wobei der Schwerpunkt dieses Abschnitts auf dem Vergleich verschiedener Erkenner und deren Eignung für das vorliegende Systemkonzept liegt. Zuletzt wird in Kapitel 7 dargestellt, wie sich auf Basis des zeitlichen Verlaufs der Instrumentennutzung direkt auf die jeweilige OP-Phase schließen lässt.

Die Umsetzung und Evaluation des in Abbildung 4.13 skizzierten übergeordneten Auswertemoduls kann aufgrund fehlender für alle Teilsysteme zusammenhängender Daten aus realen OP-Szenarien im Rahmen dieser Arbeit nicht näher betrachtet werden. Die zugehörige Diskussion erfolgt also rein konzeptionell.

5 Teilsystem: Aktivitätsanalyse innerhalb des OP-Saals

In diesem Kapitel wird das Teilsystem zur Aktivitätsanalyse in OP-Sälen näher erläutert. Hierzu werden zunächst die grundlegenden Konzepte zur Erfassung von Personen mittels Skelettextraktion und deren Bewegungen mittels Posenerkennung dargestellt. Basierend darauf werden mathematische Methoden zur Bewegungs- und Aktivitätsanalyse erarbeitet, die schließlich für weiterführende Auswertungen genutzt werden können. Außerdem untersucht dieses Kapitel, inwiefern die Re-Identifikation von Personen im medizinischen Umfeld erschwert ist und wie dem entgegengewirkt werden kann, um eine robuste Auswertung der Raumaktivität zu ermöglichen.

Die Grundidee des Teilsystems zur Aktivitätsanalyse besteht einerseits in der Erkennung einzelner konkreter Handlungen auf Basis von Bewegungsmustern verhältnismäßig kleiner Bewegungen, bspw. Bewegungen der Arme einer einzelnen Person, und andererseits der Analyse globaler Bewegungen, wie bspw. Laufwege einzelner Personen oder Änderungen der Personenanzahl im Raum. Da die einzelnen Teammitglieder während der laufenden Operation üblicherweise einen festen Platz im Raum haben, wird die Hypothese aufgestellt, dass gesteigerte Gesamtaktivität innerhalb des OP-Saals sowie Verlassen und Wiederkehren von Personen auf außerplanmäßige Ereignisse, wie die Notwendigkeit zur Besorgung zusätzlicher Materialien, hindeuten. Entsprechend kann diese Information als Anomaliedetektion genutzt werden. Dies dient nicht direkt der Phasenerkennung, kann aber Rückschlüsse auf Verzögerungen im OP-Ablauf zulassen und somit die OP-Planung präzisieren.

5.1 Systemkonzept zur Erfassung von Aktivität aus Videodaten

Die beschriebenen Analysen sollen mithilfe von Methoden der Posenerkennung umgesetzt werden. Dabei erfolgt die Detektion der Akteure und deren Körperhaltung, wie in Abschnitt 3.2 detailliert beschrieben, über die Extraktion der Skelettstrukturen durch einzelne Gelenkpunkte (sog. Keypoints) und Körperteile (Verbindungen von Keypoints).

Eine Möglichkeit zur Erkennung und Analyse von Bewegungen ist die Untersuchung der Beträge der Vektoren der zugehörigen Keypoints zwischen zwei Videoframes. Ein „Bewegungsvektor" kann also als Verbindung zwischen einem Keypoint $K(t)$ zum Zeitpunkt t mit dem zugehörigen Keypoint $K(t-1)$ zum Zeitpunkt $t-1$ betrachtet werden (siehe Abbildung 5.1).

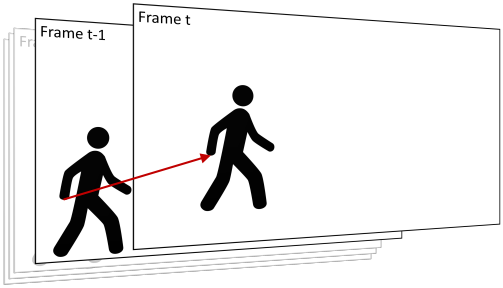


Abbildung 5.1: Bewegungsvektor (rot) zwischen zwei Videoframes

Unter Betrachtung eines Keypoints $K(t)_{p,cl}$ der Person p und der Klasse cl, wobei $p \in \mathbb{N}$ und $cl \in \{Nase, Schulter, Hand, \dots\}$, ergibt sich somit der Bewegungsvektor $\vec{m}(t)_{p,cl}$ als Differenz des gleichen Keypoints aus zwei aufeinanderfolgenden Videoframes. Gleichung (5.1) beschreibt diesen Zusammenhang formal.

$$\vec{m}(t)_{p,cl} = K(t)_{p,cl} - K(t-1)_{p,cl} \tag{5.1}$$

Zur Bestimmung der Größe einer Bewegung über einen bestimmten Zeitraum werden die Beträge der Bewegungsvektoren aus (5.1) über die vergangenen N Frames aufsummiert. Somit ergibt sich Gleichung (5.2). Die Berechnung wird für jeden neuen Frame sowie für jede Person und Klasse wiederholt, wobei $t = 0$ immer den neusten Frame bezeichnet.

$$m_{p,cl} = \frac{1}{\alpha} \sum_{t=-N+1}^{0} |\vec{m}(t)_{p,cl}| \tag{5.2}$$

Für $\alpha = N$ ergibt sich aus (5.2) gerade das arithmetische Mittel der Bewegungsvektoren. Da der Bewegungsvektor $\vec{m}$ über die Abstände zweier Bildpunkte ohne Referenz zu einer realen Längeneinheit berechnet wird, lässt sich damit zunächst keine Aussage über die absolute Größe der Bewegung treffen. Um zumindest relative Größen der Bewegungen zu erhalten, kann $m_{p,cl}$ ins Verhältnis zum Ergebnis der vorangegangenen Berechnung gesetzt werden. Dies hat aber wiederum den Nachteil, dass längere gleichmäßige Bewegungen nur schwer erkennbar sind, da sie sich lediglich zu deren Beginn oder bei schnellen Änderungen erkennen lassen. Für lange gleichmäßige Bewegungen oder wenn keine Änderung sichtbar ist, würde das Verhältnis gegen den Wert eins konvergieren.

Des Weiteren ist es möglich die Berechnung an unterschiedliche Bewegungsarten anzupassen. Durch Anpassung von N, also des Zeitraumes, über den in (5.2) aufsummiert wird, kann die Analyse der Bewegungsgröße beeinflusst werden. Dabei können für kleine N, also kurze Auswertezyklen, tendenziell kleinere und schnellere Bewegungen und entsprechend für große N größere und länger andauernde Bewegungen analysiert werden.

Abbildung 5.2 zeigt den konzipierten Ablauf für die Bewegungsanalyse. Da für die Berechnung der Bewegungsvektoren immer die Daten von zwei aufeinanderfolgenden Frames notwendig sind, speichert das System immer die Keypoints der beiden zuletzt ausgewerteten Bilder. Dies ist in Abbildung 5.2 durch ein Schieberegister der Länge 2 dargestellt. Sobald neue Keypoints geladen werden, werden diese in das Schieberegister übergeben und die älteren der zwei

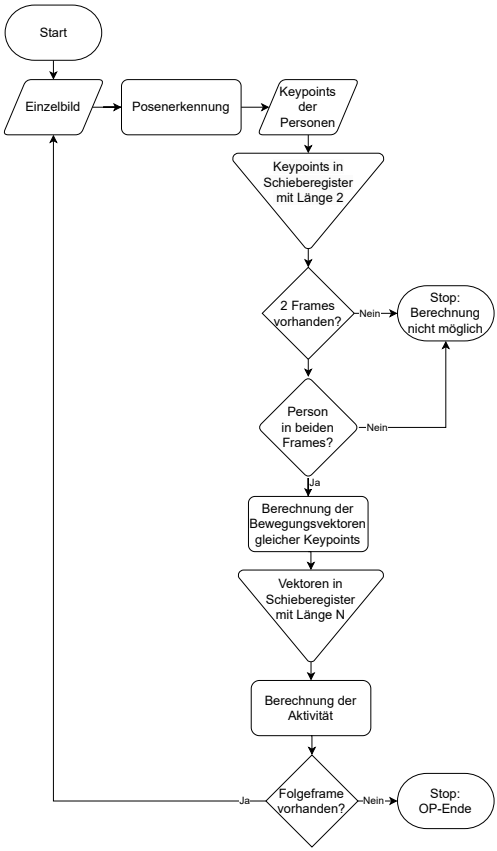


Abbildung 5.2: Flussdiagramm zur Aktivitätsanalyse

gespeicherten Keypoints werden verworfen. So stehen immer die Keypoints der letzten zwei Frames für die weiteren Verarbeitung zur Verfügung. Bevor die Bewegungsvektoren berechnet werden können, muss geprüft werden, ob die detektierten Personen aus dem aktuellen Frame auch im vergangenen Frame sichtbar waren. Ist das nicht der Fall, weil beispielsweise eine Person den Bildbereich verlassen hat, so kann für diese Person kein Bewegungsvektor berechnet werden. In Abbildung 5.2 ist dies durch einen Desicion-Block und gegebenenfalls durch den Abbruch der Berechnung dargestellt. Sind die Personen in beiden

Frames vorhanden, kann die Berechnung der Bewegungsvektoren durchgeführt werden. Dazu wird, wie in Gleichung (5.1) definiert, der Vektor zwischen zwei zueinander gehörenden Keypoints berechnet. Dargestellt wird die Berechnung in Abbildung 5.2 durch einen Process-Block. Die Ergebnisse der Berechnung werden an ein weiteres Schieberegister übergeben. Das Schieberegister besitzt N Speicherplätze, je nach Wahl von N in Gleichung (5.2). Die N Bewegungsvektoren können dann an die Auswertemetrik gemäß Gleichung (5.2) übergeben werden. Nach der Berechnung der Aktivität werden neue Keypoints ausgelesen und der Ablauf wiederholt sich.

5.2 Umsetzung und Auswertung der Bewegungsanalyse

Die beschriebene Auswertemetrik wurde in Python auf System 1 aus Tabelle A.1 im Anhang umgesetzt. Alle Ergebnisse der Posenerkennung werden in „JavaScript Object Notation (JSON)“-Notation gepeichert. Abbildung 5.3 zeigt einen Ausschnitt einer solchen JSON-Datei mit den Ergebnissen des Posentrackings für eine Person. JSON nutzt Key-Value-Paare zur Repräsentation der enthaltenen Daten. Folgende Keys sind nach [46] in Abbildung 5.3 dargestellt:

- image id: Nummer des Videoframes.
- category id: Nummer, welche einer Objektkategorie (hier Personen-ID) zugeordnet ist.
- keypoints: Koordinaten und Detektionssicherheiten der detektierten Keypoints.
- score: Detektionssicherheiten der ganzen Person.
- box: Eckpunkte einer Boundary Box, welche die Person umgibt.
- idx: ID um Personen zuzuordnen.

Die Daten aus Abbildung 5.3 sind nur für eine Person und einen einzelnen Frame dargestellt, liegen aber für alle detektierten Personen sowie für jeden Videoframe vor. Die Keypoints liegen in der Reihenfolge $x_0, y_0, c_0, x_1, y_1, c_1, \ldots$ vor, wobei x und y die Position des Keypoints und $c \in [0; 1]$ die Detektionssicherheit beschreiben [46]. Es beschreiben also immer drei Werte einen bestimmten Keypoint und die darauffolgenden drei den Nächsten. Die Reihenfolge, in welcher die Keypoints angeordnet werden, ist dabei bekannt.


```
[
  {
    "image_id": "0.jpg",
    "category_id": 1,
    "keypoints": [
      322.1070861816406,
      26.123435974121094,
      0.948883056640625,
      327.4691467285156,
      18.08030891418457,
      0.9360940456390381,
      314.06396484375,
      18.08030891418457,
      0.9499721527099609,
      ...,
      ...,
      ...,
    ],
    "score": 3.0266048908233643,
    "box": [
      247.69993591308594,
      0.11733205616474152,
      167.5815887451172,
      274.5386962890625
    ],
    "idx": 1
  },

  {...},
  {...},
  ...
]
```


Abbildung 5.3: Ausschnitt einer JSON-Datei mit Rohdaten, generiert durch Pose Tracking

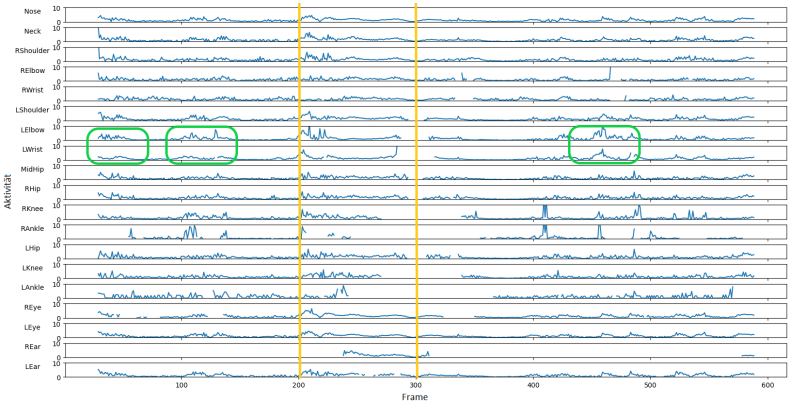
Im Rahmen der Implementierungsarbeiten stellte sich heraus, dass eine Überwachung der aufgezeichneten Keypoints bzw. der daraus resultierenden Bewegungsvektoren und das Festlegen von α auf einen festen Wert, z.B auf deren Maximum oder Mittel bessere Ergebnisse liefert. Ein Vorteil dieses Vorgehens ist, die unabhängig von vergangenen Werten dargestellte Bewegung, die eine Verfälschung durch vorherige besonders hohe oder niedrige Aktivität verhindert.

Diese Kalibrierung muss bei der Einrichtung des Systems einmalig durchgeführt werden oder kann alternativ in regelmäßigen Abständen automatisiert erfolgen. Es zeigte sich, dass bei einer Framerate von $F = 30fps$ und $N < \frac{F}{fps} = 30$ in (5.2) besonders kleine und schnelle Bewegungen (z.B. Armbewegungen) erkannt werden können. Für $N = k \cdot \frac{F}{fps}$ mit $k = 1, 2, 3, ..$ sind hingegen besonders große oder lange Bewegungen gut erkennbar, da die jeweilige Bewegung über einen längeren Zeitraum betrachtet wird.

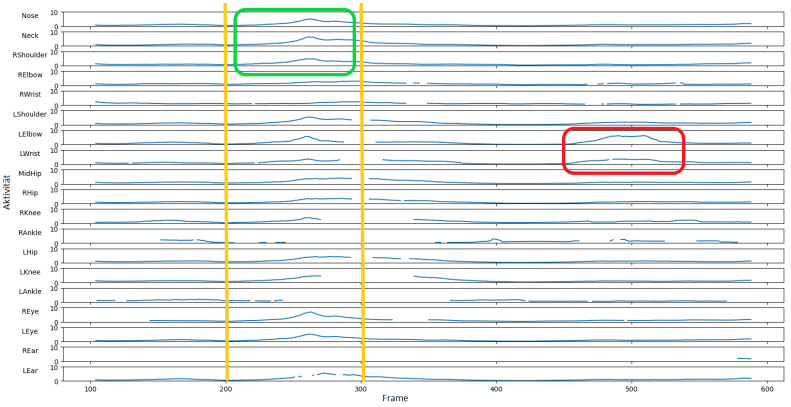
In Abbildung 5.4 sind zwei Darstellungen derselben Bewegung abgebildet. Dargestellt ist jeweils das Ergebnis aus Gleichung (5.2) für eine unterschiedliche Anzahl N an Summanden. Für α wurde ein fester Wert gewählt. Zwischen den gelben Markierungen findet im Testvideo eine große Bewegung statt. In Abbildung 5.4b ist die Bewegung deutlich zu erkennen, wohingegen sie in Abbildung 5.4a nicht eindeutig ersichtlich ist. Da der Bewegungsvektor keine reale Größe ist, bleibt auch dessen Summe einheitenlos. Sie kann jedoch als Vielfaches von $\frac{1}{\alpha}$ betrachten werden, wodurch sie vergleichbar wird.

Bei der Datenauswertung stellte sich heraus, dass die Bewegungsdaten verschiedener Personen untereinander eine hohe Korrelation aufweisen. Deshalb wurde untersucht, unter welchen Umständen dieser Effekt auftritt und wie er behoben werden kann. Bei genauerer Betrachtung der Keypoints stellte sich heraus, dass die Zuordnung der Personen nicht einheitlich ist. In Abbildung 5.5 sind fünf aufeinanderfolgende Frames abgebildet, wobei jede Farbe eine Person repräsentiert. Hierbei wird ersichtlich, dass sich die Farben einzelner Personen von Frame zu Frame ändern, was bedeutet, dass die detektierten Keypoints unterschiedlichen Personen zugeordnet wurden. Durch diese Vermischung von Personen werden auch deren Bewegungen bei der Auswertung vermischt. Das bedeutet, dass die Bewegungsdaten einer Person in Wirklichkeit aus einer Kombination der Bewegungen von unterschiedlichen Personen bestehen. Dies erklärt auch die hohe Ähnlichkeit der Bewegungsdaten verschiedener Personen.

Um konsistente Bewegungsdaten zu erhalten, ist es entsprechend notwendig einzelne Personen über einen längeren Zeitraum zu verfolgen und auf diese Weise eine Vermischung von Bewegungen zu vermeiden. Durch dieses Tracking



(a) Auswertung für kleine Bewegungen, $N = 1$

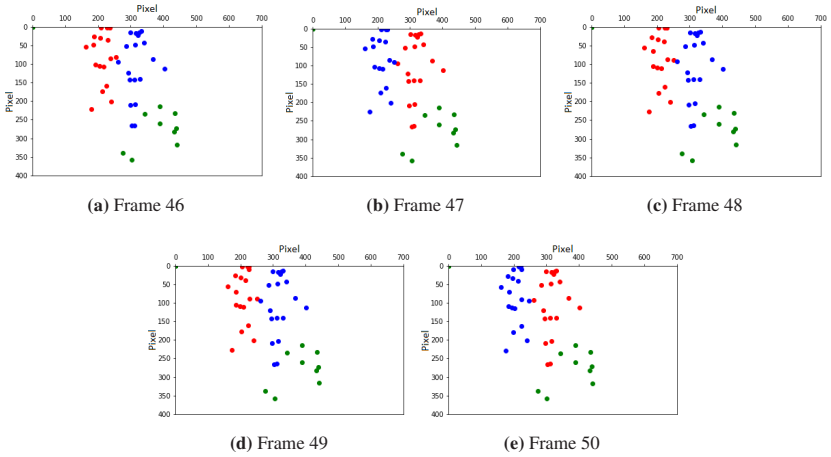


(b) Auswertung für große Bewegungen, $N = 60$

Abbildung 5.4: Auswirkung unterschiedlicher Auswertungen. Zwischen den gelben Markierungen fand eine Bewegung des gesamten Körpers („große Bewegung“) statt.

können die Keypoints dann immer der selben Person zugeordnet werden. Dieses Verfahren bezeichnet man als „Posentracking“ (vgl. Abschnitt 3.2). Die Bibliothek „AlphaPose“ bietet neben der Posenerkennung auch die Möglichkeit des

Posentrackings, weshalb im weiteren Vorgehen mit dieser Bibliothek gearbeitet wird. Durch das Tracking wird nun die korrekte Zuordnung der Keypoints sichergestellt, sodass das zuvor beschriebene Problem vermieden wird.



(a) Frame 46 **(b)** Frame 47 **(c)** Frame 48

(d) Frame 49 **(e)** Frame 50

Abbildung 5.5: Zuordnungsproblem der Personen zwischen Frames

Zur besseren Veranschaulichung der Ergebnisse wurden die Daten grafisch dargestellt. In Abbildung 5.6 sieht man die Ergebnisse einer Bewegungsdetektion, welche durch Posenerkennung, Posentracking und die anschließende Datenverarbeitung entstanden sind. In Y-Richtung sind die Frames aufgetragen und in X-Richtung das Ergebnis aus (5.2), gemittelt über alle Keypoints und für jede Person. Da Personen, welche das Bild verlassen und anschließend erneut zu sehen sind, immer eine neue ID zugeordnet wird, können in den Daten mehr Personen aufgeführt sein, als tatsächlich vorhanden waren. Ebenso können durch Fehldetektionen zusätzliche Personen in den Daten entstehen, die im weiteren Verlauf allerdings keine Aktivität aufweisen. In Abbildung 5.6 zeigt sich dies dadurch, dass für einzelne Personen, über den kompletten Zeitraum der Detektion keine Bewegung in den Daten zu erkennen ist.

Durch Re-ID von Personen kann dieses Verhalten vermieden werden, was jedoch kein triviales Problem darstellt.

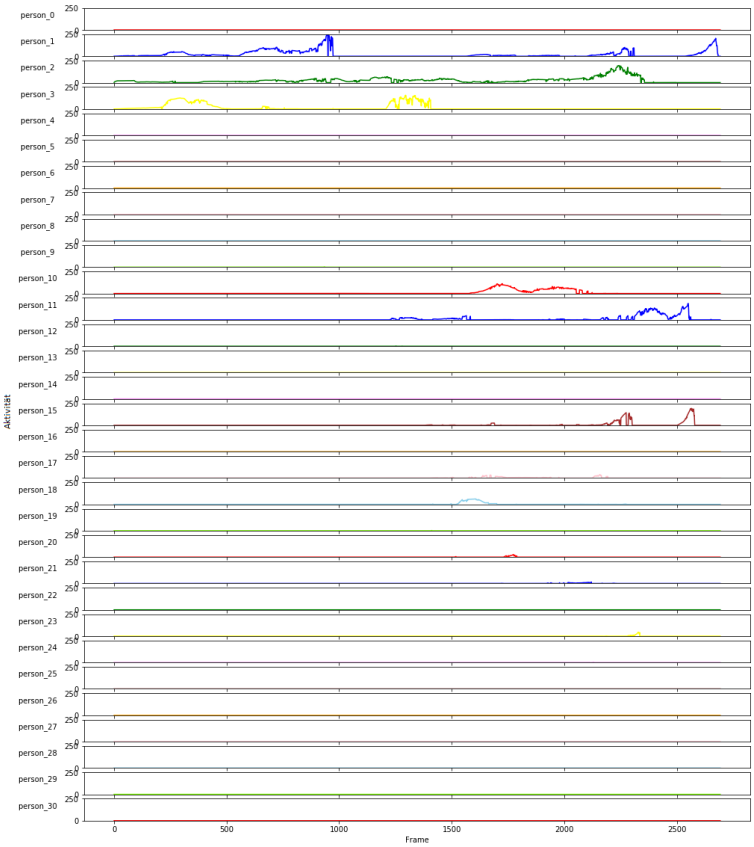


Abbildung 5.6: Aktivitätsdetektion für mehrere Personen

5.3 Effekte medizinischer Kleidung auf Algorithmen zur Re-Identifikation von Personen

In den folgenden Abschnitten wird wegen des zuvor beschriebenen Problems der fehlenden Zuordnung von bestehenden IDs zu zuvor bereits erkannten Personen nach deren Wiedereintritt ins Kamerabild (Re-ID) untersucht, inwiefern das

Problem im medizinischen Umfeld begünstigt wird und wie dieser Problematik begegnet werden kann.

Da annotierte Daten in ausreichender Qualität und Quantität, speziell im medizinischen Umfeld, nur selten verfügbar und schwer zu erzeugen sind, analysiert das nachfolgend beschriebene Experiment außerdem, ob bestehende Algorithmen zur Re-ID von Personen in Alltagskleidung aus Szenarien, in denen weniger markante, einheitliche Kleidung getragen wird, trotzdem funktionieren. Insbesondere in OP-Sälen, tragen die Mitarbeitenden nicht nur einheitliche Kleidung, sondern viele ihrer individuellen Merkmale, wie Frisuren oder Gesichtszüge, werden z. B. durch einen Mund-Nasen-Schutz verdeckt. Aus diesem Grund wird mithilfe dieses Experiments konkret untersucht, wie sich medizinische Kleidung auf die Re-Identifikationsleistung verschiedener bestehender und mit Daten aus Alltagsszenarien trainierter Modelle auswirkt. Weiterhin wird betrachtet, wie der Re-Identifikationssprozess mit individuellen Markern verbessert werden kann. Marker sind eindeutige Identifizierungsmerkmale, wie bspw. Nummern, QR-Codes oder individuelle Kleidungsstücke. Die Erkenntnisse hieraus schaffen bessere Voraussetzungen für die Erarbeitung neuartiger Konzepte zur Handlungserkennung im medizinischen Bereich.

Die Re-Identifizierung von Personen, die stark verdeckt sind oder sehr ähnlich aussehen, wird in bisherigen Arbeiten nur selten behandelt. Diese Untersuchung bildet somit einen Beitrag zur Schließung dieser Lücke und erzeugt wertvolle Erkenntnisse, wie bestehende Deep-Learning-Modelle in der Lage sind, aus Trainingsdaten zu generalisieren und ob sie direkt in Umgebungen eingesetzt werden können, die sich deutlich von den Anwendungen aus dem ursprünglichen Trainingsprozess unterscheiden. Um die Auswirkungen medizinischer Kleidung zu untersuchen, werden fünf verschiedene bestehende Re-ID-Modelle auf einem nicht-medizinischen Datensatz trainiert und dann mit Bildern von Personen in medizinischer Kleidung getestet.

Die Inhalte dieses Abschnitts basieren überwiegend auf der in [4] bereits zuvor publizierten Arbeit.

5.3.1 Messsetup & Probandenkollektiv für die Datenakquise

Da zum Zeitpunkt der Erstellung dieser Arbeit kein Datensatz für die medizinische Re-ID zur Verfügung steht, ist zunächst die Aufzeichnung und Annotation geeigneter Daten für die beschriebene Untersuchung erforderlich. Zusätzlich zur medizinischen Kleidung sollen auch spezifische Marker an den Personen enthalten sein, um deren Auswirkung auf die Re-Identifikation analysieren zu können. Zum Ausschluss personenspezifischer Merkmale sollen Aufnahmen der selben Personengruppe in ihrer Alltagskleidung als Referenz verwendet werden. Somit sind die gleichen Personen in drei unterschiedlichen Erscheinungsbildern im Datensatz enthalten.

Zur Erstellung der Bilddaten für den Datensatz wurde eine Gruppe von Personen, die Tätigkeiten in einer medizinischen Umgebung nachbilden, über eine Deckenkamera aufgenommen. Insgesamt beteiligten sich dabei sechs Probanden. Eine Beschreibung der Teilnehmer ist in Tabelle A.3 im Anhang zu finden. Um die zuvor diskutierten Forschungsprobleme zu untersuchen, wurden die Aufnahmen insgesamt drei Mal mit den gleichen Probanden in unterschiedlicher Kleidung durchgeführt. Die Laufzeit jedes Durchganges beträgt dabei jeweils 15 Minuten. Die erste Aufnahme (OR0) zeigt Personen in ihrer normalen Kleidung bei der Ausführung der Aufgaben. Aufgrund der zum Aufnahmezeitpunkt vorherrschenden Covid-19-Pandemie war es auch hierbei notwendig, einen Mund-Nasen-Schutz zu tragen, obwohl dies konzeptionell eigentlich nicht vorgesehen war. Dieser Datensatz dient als Referenz für die geplanten Untersuchungen. Im zweiten Datensatz (OR1) tragen die Personen einheitliche medizinische Kleidung. Diese besteht jeweils aus einer OP-Haube, einem Mund-Nasen-Schutz, einem OP-Kittel, Handschuhen, sowie Schuhüberziehern. Im dritten Datensatz (OR2) wurde die medizinische Kleidung der Personen mit individuellen Markierungen versehen, um die Re-ID zu erleichtern. Zur Umsetzung der individuellen Markierungen wurden verschiedenfarbige OP-Hauben verwendet, sodass jeder Person eine individuelle Farbe zugewiesen wurde (siehe Abb. 5.7). OP-Hauben

sind eine geeignete Wahl, da sie im medizinischen Umfeld ohnehin verwendet werden und auch aus der Perspektive der Deckenkamera gut sichtbar sind. Während jeder Aufnahme führten die Probanden verschiedene, den Handlungen im medizinischen Umfeld angelehnte Aufgaben aus, die an acht verschiedenen Stationen im Raum verteilt waren (vgl. Abb. 5.8). Zu den Aufgaben gehörten die Arbeit mit medizinischen Instrumenten, das Einsortieren von Tabletten in Behälter oder das Dokumentieren von Vorgängen mit einem Tablet-Computer. Die einzelnen Stationen sind in Tabelle A.4 im Anhang genauer erläutert. Sobald eine Aufgabe abgeschlossen war, wechselten die Probanden, entsprechend der Nummerierung in Abbildung 5.8, zur nächsten Station an einem anderen Ort im Raum und begannen sofort mit der nächsten Aufgabe. Einige Stationen befanden sich außerhalb des Sichtbereichs der Kamera, um die Abwesenheit von Personen sowie deren späteres Wiederauftauchen zu simulieren. Dies ist für den Aufbau des Datensatzes in dieser Untersuchung nicht direkt relevant, da hierfür nur Bilder von einzelnen Personen notwendig sind. Es dient jedoch als Grundlage für künftige Tests von Echtzeit-Wiedererkennungssystemen auf Videodaten. Der Grundgedanke hinter der Durchführung verschiedener Aufgaben während der Aufnahme ist nicht die Schaffung einer realistischen Umgebung, sondern die Erhöhung der Intra-Klassen-Variation (vgl. Abschnitt 2.2.3.3) in den Bildern der Einzelpersonen. Diese wird durch Bewegungen der Personen bei der Aufgabenausführung oder der Arbeit mit medizinischen Geräten erzielt. Darüber hinaus kommt es bei Stationswechseln zu vermehrter Verdeckung zwischen den Personen, was realitätsnahe Analysen weiter begünstigt. Abb. 5.9 gibt einen Eindruck über den gewählten Aufbau im Raum während der Aufnahmen. Die Aufgaben und die Reihenfolge, in der sie erledigt wurden, blieben bei allen drei Aufnahmen gleich.

Um die entsprechenden Datensätze aus den einzelnen Aufnahmen zu erstellen, wurden ausgewählte Einzelbilder des Videos mit Annotationen versehen. Da die verschiedenen Aufgaben keine große körperliche Aktivität erforderten, war die Gesamtbewegung der Probanden sehr gering, sodass sich zwischen den Einzelbildern nur wenig änderte. Zur Vermeidung stark identischer Bilder im Datensatz wurde nur jeder 100. Frame annotiert. Zu den Annotationen gehören

Abbildung 5.7: Stichproben aus dem OR2-Datensatz. Zur Verbesserung der Wiedererkennungsleistung ist jede Person durch eine individuelle farbige OP-Haube markiert.

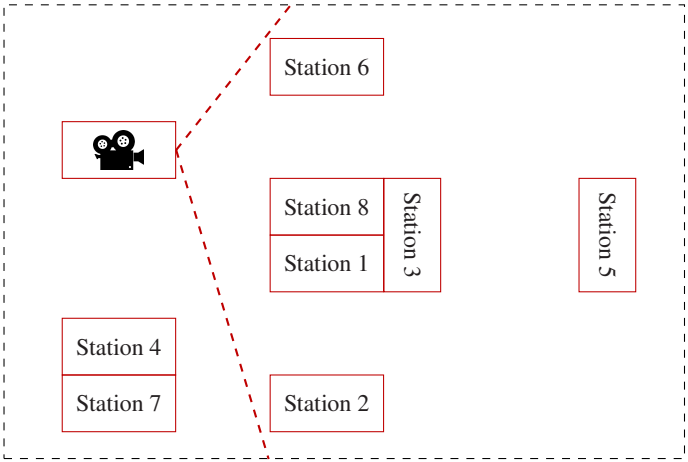
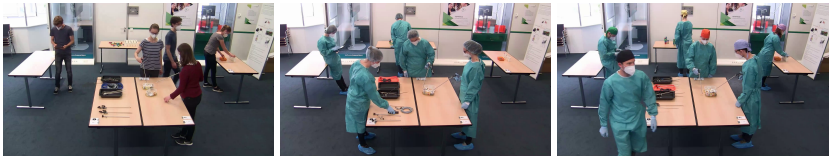


Abbildung 5.8: Verteilung der Stationen im Raum. Während der Aufzeichnung wechseln die Personen zwischen den Stationen. Der Sichtbereich der Kamera ist durch eine gestrichelte rote Linie markiert. Die Stationen 4 und 7 befinden sich außerhalb des Sichtbereiches, um eine längere Abwesenheit von Personen zu simulieren.

die Bounding Box für die Lokalisierung im Bild und eine eindeutige ID für jede Person. Die IDs sind zwischen den Einzelbildern und zwischen allen drei Aufnahmen konsistent. Alle Annotationen wurden sorgfältig von Hand gesetzt, um deren Korrektheit sicherzustellen. Aus den annotierten Frames wurden die Bilder der einzelnen Personen ausgeschnitten und in den entsprechenden Datensatz aufgenommen. Auf diese Weise wurden etwa 3.600 annotierte Bilder erzeugt, die sich gleichmäßig auf die Datensätze OR0, OR1 und OR2 verteilen.



(a) OR0 **(b)** OR1 **(c)** OR2

Abbildung 5.9: Setting während der Aufnahmen der Datensätze OR0 mit normaler Kleidung (links), OR1 mit medizinischer Kleidung (Mitte) und OR2 mit einzelnen farbigen OP-Hauben (rechts). Während das Setting und die Aufgaben bei jeder Aufnahme gleich bleiben, wechselt zur Untersuchung deren Beeinflussung die Kleidung der Personen.

5.3.2 Modellauswahl, Training & Inferenz

Der Grundgedanke dieser Untersuchung besteht nicht darin, ein neues Modell für die medizinische Re-Identifizierung zu schaffen, sondern die Auswirkungen der medizinischen Kleidung auf verschiedene bestehende Modelle zu analysieren. Aus diesem Grund wurde das Training der Modelle nicht mit medizinischen Daten durchgeführt, sondern der in der Forschung zur Re-ID von Personen weit verbreitete Market1501-Datensatz [171] verwendet (vgl. Abschnitt 3.4.1). Mit der vorhandenen Menge an Bildern eignet sich der Market1501-Datensatz für das Modelltraining besser als der selbst aufgenommene OR0-Datensatz.

Im Rahmen der Analyse wurden insgesamt fünf verschiedene Modelle untersucht. Für einen objektiven Vergleich der Architekturen, unabhängig von Implementierungsdetails wie z. B. Hyperparameter oder Datensatzkonfiguration, wurde eine Programmierbibliothek verwendet, die es erlaubt, den Modellkern von der Implementierungsumgebung und anderen Parametern zu trennen. Damit konnte für jedes Modell dieselbe Datensatzkonfiguration und dieselben Parameter verwendet werden. Für die Implementierung der Modelle wurde entsprechend die Bibliothek Torchreid [173] verwendet.

Für das Experiment wurden folgende Modelle ausgewählt:

1. OSNet [174]
2. OSNet-AIN [175]

3. PCB [146]
4. ResNet-Mid [169]
5. MLFN [26].

Von allen Modellen, die zum Zeitpunkt des Experiments mit Torchreid verfügbar waren, wurden diejenigen ausgewählt, die speziell für die Aufgabe der Re-Identifikation von Personen geeignet sind. Alle Modelle wurden mit dem Cross-Entropy Loss als Kostenfunktion und der euklidischen Distanz als Matching-Distanzmetrik implementiert. Eine Übersicht der relevantesten Parameter bietet Tabelle 5.1. Eine detaillierte Beschreibung der dort genannten Parameter findet sich in der Dokumentation zu Torchreid [173].

Tabelle 5.1: Übersicht verschiedener Parameter für das Training und die Validierung der Modelle. Die Bedeutung der Parameter findet sich in der Dokumentation von Torchreid [173]

	Parameter	Wert
Data	Height	256
	Width	128
	Combineall	False
	Augmentierung	Random Flip
Loss	Name	Softmax
	Label-Smooth	True
Train	Optimizer	Amsgrad
	Learning Rate	0,0015
	LR-Scheduler	Cosinus
	Epochen	100
	Batch-Size	64
Test	Batch-Size	300
	Distanzmetrik	euklidisch

Nach dem Training der Modelle, wurden sie mit den zuvor beschriebenen Datensätzen OR0, OR1 und OR2 getestet. Der Test mit dem OR0-Datensatz diente als Grundlage für spätere Vergleiche, da die Bilder im OR0-Datensatz,

ähnlich wie im Trainingsdatensatz, ebenfalls Personen in Alltagskleidung zeigen. Da in OR1 dieselben Personen in derselben Umgebung wie in OR0 abgebildet sind, zeigt der Vergleich der Ergebnisse den direkten Einfluss der medizinischen Kleidung auf die Leistung der Modelle. Die Ergebnisse von OR2 können dann in Bezug auf OR1 betrachtet werden und zeigen die Auswirkungen der zusätzlichen Markierungen durch farbige OP-Hauben. Abbildung 5.10 gibt einen Überblick darüber, wie die Datensätze mit den Modellen zusammenhängen und welche Datensätze für das Training und die Inferenz verwendet wurden. Durch den Vergleich aller Testergebnisse lassen sich die Auswirkungen von medizinischer Kleidung, farbigen OP-Hauben sowie Unterschiede zwischen den Modellen feststellen.

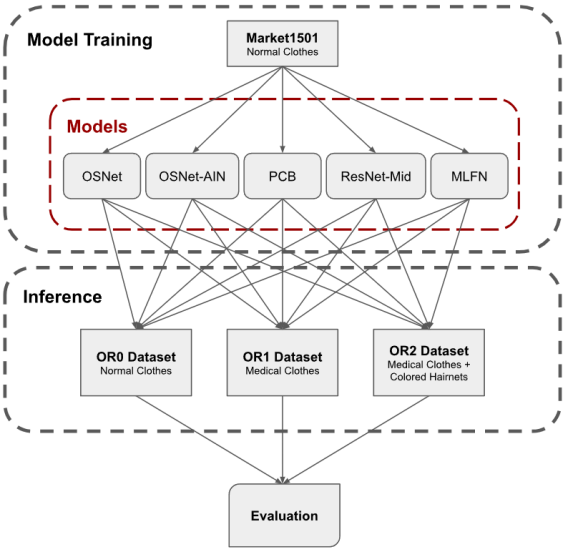


Abbildung 5.10: Gesamtpipeline des Modelltrainings und der Inferenz. Die Abbildung zeigt den Zusammenhang zwischen den Modellen und den für das Modelltraining und die Inferenz verwendeten Datensätzen.

Um einen Vergleich zum häufig in der Praxis eingesetzten Finetuning von Modellen mit einem beschränkt großen Datensatz herstellen zu können, wurde

zusätzlich untersucht, wie sich das Finetuning der gewählten Modelle mit den OR1-Daten auf das Ergebnis auswirkt. Dafür wurden die Modelle nach dem Training mit Market1501 mit einem Teil der Daten aus OR1 noch einmal verfeinert. Abbildung 5.11 zeigt entsprechend die Zusammenhänge zwischen Modellen, Daten und Inferenz.

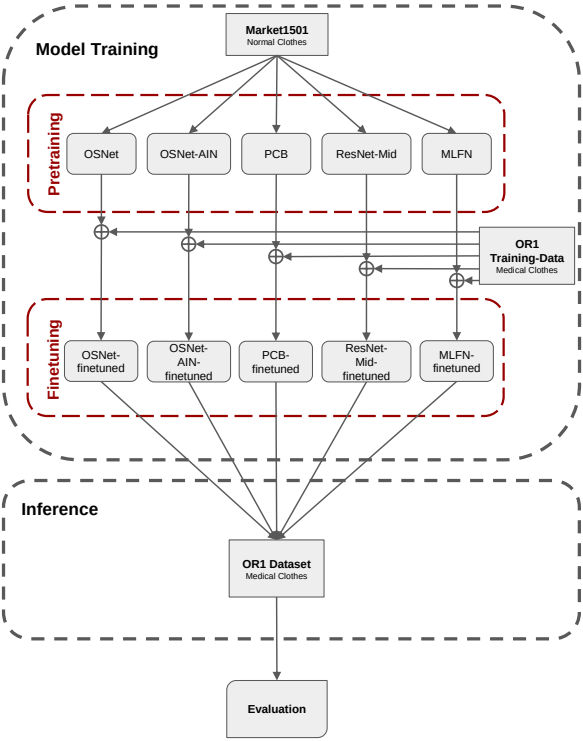


Abbildung 5.11: Gesamtpipeline des Modelltrainings und der Inferenz mit Finetuning. Die Abbildung zeigt den Zusammenhang zwischen den Modellen und den für das Modelltraining und die Inferenz verwendeten Datensätzen.

Die Einhaltung etablierter Bewertungsmetriken gewährleistet die Vergleichbarkeit der Ergebnisse mit früheren und zukünftigen Arbeiten. Daher wurden die

mAP und die Rang-1 Genauigkeit für die Modellevaluation verwendet, die üblicherweise bei der Re-ID von Personen eingesetzt werden (vgl. [166], [50],[165], [168]). Auf diese Weise lassen sich die Ergebnisse objektiv beurteilen und mit verwandten Arbeiten vergleichen.

Die Aufteilung des Datensatzes in Queries und Galeriebilder (sog. „Split“) wurde für alle verwendeten Datensätze auf ein Verhältnis von 1:5 festgelegt. Die Tabelle in Anhang A.5 enthält zusätzliche Informationen zu den Splits der einzelnen Datensets für die Untersuchungen. Hervorzuheben ist hierbei insb. der Split zum Finetuning der Modelle mit OR1 (vgl. Tabelle A.5). Da ein Teil der Daten für das Training benötigt wurde, wurde sowohl die Query als auch die Galerie erheblich verkleinert, was möglicherweise Auswirkungen auf die Aussagekraft der Ergebnisse hat. Dies wird in der Ergebnisdiskussion genauer betrachtet.

5.3.3 Auswertung der Effekte medizinischer Kleidung auf die Re-Identifikationsqualität

Die Ergebnisse des Experiments sind in Tabelle 5.2 dargestellt. Darin ist für alle getesteten Modelle eine Abnahme der Rang-1 Genauigkeit zwischen den OR0- und den OR1-Datensätzen ersichtlich. Der größte Einbruch tritt hierbei bei der Verwendung des Modells ResNet-Mid mit einer Differenz von 43 Prozentpunkten auf. Die geringste Differenz zwischen OR0 und OR1 ergibt sich für OSNet-AIN mit einer Abweichung von 13 Prozentpunkten.

Zwischen OR1 und OR2 ist ein allgemeiner Anstieg der Rank-1 Genauigkeit zu erkennen. Der größte Zuwachs mit 30,7 Prozentpunkten liegt bei ResNet-Mid, der kleinste bei OSNet-AIN mit 8,7 Prozentpunkten. Die geringe Steigerung für OR2 bei OSNet-AIN liegt darin begründet, dass für OR1 die beste Leistung erreicht werden konnte und dementsprechend auch ein geringeres Steigerungspotenzial vorhanden war. Über alle Datensätze hinweg erzielt OSNet-AIN die höchste Rank-1-Genauigkeit. Insgesamt nähern sich die Ergebnisse des OR2-Datensatzes zu denen von OR0 an, fallen aber stets schlechter aus.

Tabelle 5.2: Rang-1 Genauigkeit und mAP verschiedener Modelle für die Datensets OR0, OR1 und OR2.

	Rank-1 (%)			mAP (%)		
Modell	OR0	OR1	OR2	OR0	OR1	OR2
OSNet	99.5	69.0	92.9	79.5	25.2	32.8
OSNet-AIN	**100.0**	**87.0**	**95.7**	**80.2**	**27.5**	**37.1**
PCB	98.5	67.0	88.6	65.0	23.7	29.5
ResNet-Mid	97.5	54.5	85.2	57.7	23.0	31.1
MLFN	95.1	67.0	90.0	52.1	22.1	30.8

Die Ergebnisse in Tabelle 5.2 zeigen auch eine Abnahme der mAP zwischen den OR0- und OR1-Datensätzen. Dies ist wiederum bei allen getesteten Modellen zu beobachten. Der größte Rückgang von 54,3 Prozentpunkten tritt bei Verwendung von OSNet auf, der geringste mit 30 Prozentpunkten bei Verwendung von MLFN. Wie bereits im Zusammenhang mit der Rang-1 Genauigkeit festgestellt, erreicht OSNet-AIN auch die höchste mAP für alle Datensätze. Insgesamt ist ein geringer Anstieg der mAP von OR1 zu OR2 zu erkennen. Dennoch liegt die mAP immer noch deutlich unter der von OR0.

Im Anhang A.2.2 finden sich zusätzliche Grafiken zur genaueren Ergebnisanalyse anhand der Rank-1 Genauigkeit und der mAP für die untersuchten Modelle. Darüber hinaus befinden sich dort jeweils drei Konfusionsmatrizen, entsprechend der drei Datensets OR0, OR1 und OR2, zu jedem der getesteten Re-ID-Modelle. Diese dienen als Grundlage für die Berechnung der genannten Ergebnisse.

Abbildung 5.12 stellt die Ergebnisse nach Finetuning der Modelle dar. Wie zu sehen ist, kann hierdurch sowohl die Rang-1 Genauigkeit als auch die mAP verbessert werden. Auffällig dabei ist, dass sich die Ergebnisse ohne Finetuning gegenüber der zuvor gewählten Aufteilung der Daten (Abbildung A.11b) verschlechterten. Dies ist vor allem dadurch zu begründen, dass für den im Finetuning verwendeten Split und durch die limitierte Größe des Datensets,

die Galerie sehr klein wird und weniger Queries überprüft werden können. Diese Kombination führt dazu, dass weniger korrekte Zuordnungen zwischen den Queries und der Galerie gefunden werden. Des Weiteren zeigt sich, dass sich die mAP mit Finetuning in Abbildung 5.12 nicht nur gegenüber der mAP ohne Finetuning verbessert hat, sondern auch die Werte aus Abbildung A.11b übertrifft. Beispielsweise kann die mAP des OSNets im Vergleich zu Abbildung A.11b durch Finetuning mehr als verdoppelt werden. Das deutet darauf hin, dass das Finetuning einen größeren Einfluss auf die mAP hat als die Größe der Galerie. Folglich kann das Finetuning der Modelle dazu beitragen, die Rang-1 Genauigkeit und vor allem die mAP zu erhöhen.

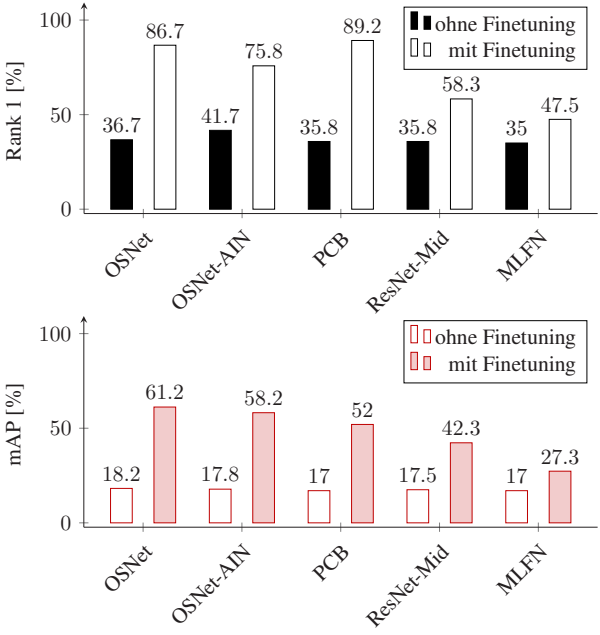


Abbildung 5.12: Auswirkungen des Finetunings für den OR1-Datensatz

5.3.4 Ergebnisdiskussion zum Einfluss medizinischer Kleidung auf die Re-Identifikation von Personen

Wie die Ergebnisse zeigen, führt die medizinische Kleidung zu einem allgemeinen Rückgang der Rang-1 Genauigkeit sowie der mAP. Dies war zu erwarten, da die medizinische Kleidung große Teile des Körpers der Probanden und damit potenziell hilfreiche Unterscheidungsmerkmale verdeckt. Das Hinzufügen von farbigen OP-Hauben, die als individuelle Marker fungieren, verbessert die Leistung der Modelle im Vergleich zu den Ergebnissen ohne farbige OP-Haube. Aus menschlicher Sicht ist das sinnig, da jede Person eindeutig identifiziert werden kann, solange die OP-Haube sichtbar und die Farbe nicht verfälscht ist. Die farbigen OP-Hauben stellen ein Unterscheidungsmerkmal dar, welches von den Modellen während des Re-Identifikationsprozesses genutzt werden kann. Es ist anzumerken, dass die Ergebnisse von OR2, entgegen der Erwartung, nicht über denen liegen, die mit Alltagskleidung erzielt werden. Nach menschlichem Ermessen sind die Personen jedoch eindeutig durch die farbigen OP-Hauben unterscheidbar. Der naheliegendste Grund dafür ist, dass die Modelle nicht mit dem OR2-Datensatz trainiert wurden und daher nie gelernt haben, dass die Farbe der OP-Hauben für die Re-Identifikation eine wichtigere Rolle spielt als andere Merkmale. Obwohl die Rang-1 Genauigkeit mit dem Hinzufügen der farbigen OP-Hauben deutlich ansteigt, steigt die mAP nicht im gleichen Maße und bleibt deutlich unter der OR0-Referenz. Dies bestätigt die vorherige Aussage, da die Modelle zwar über ein Merkmal zur Unterscheidung von Personen verfügen, aber aufgrund des fehlenden Trainings mit den spezifischen Daten mit visuellen Markern und der entsprechenden Labels nicht in der Lage sind, zu generalisieren.

Auffällig ist, dass OSNet-AIN die höchste Rang-1 Genauigkeit und die höchste mAP für alle drei Datensets erreicht. Der Grund dafür könnte die von OSNet-AIN verwendete *instance normalization* sein, die eine Normalisierung der Bilder anhand des Mittelwerts und der Standardabweichung einer Stichprobe vornimmt. Dadurch können datensatzspezifische Stile, die z. B. durch unterschiedliche Umgebungen und Lichtverhältnisse verursacht werden, eliminiert werden, was zu

stärkerer Generalisierung führt. Auch wenn die *instance normalization* die Leistung verbessert, können bessere Ergebnisse nicht allein darauf zurückgeführt werden. OSNet, das die gleiche Architektur verwendet wie OSNet-AIN, allerdings ohne *instance normalization*, erzielt dennoch bessere Ergebnisse als die übrigen Modelle. Zhou et al. verwenden *omni-scale features*, die eine Mischung von Merkmalen aus verschiedenen Skalen enthalten. Sie argumentieren, dass Merkmale kleiner lokaler Regionen ebenso wichtig sind wie globale Merkmale des gesamten Körpers. Wie die Ergebnisse zeigen, erweist sich diese Methode bei den untersuchten Datensätzen als effektiv. Da das Erscheinungsbild der Individuen nahezu identisch ist, hilft die Fokussierung auf unterschiedlich skalierte Merkmale bei deren Identifizierung. Es liegt nahe, dass die Fokussierung nicht nur auf das Gesamterscheinungsbild, sondern auch auf mehrere kleinere Merkmale oder Details zu einer besseren Re-ID führt.

Bei der Inferenz mit dem OR2-Datensatz treten Fehler, mit Ausnahme des Modells ResNet-Mid, nur vereinzelt auf. Daraus lässt sich ableiten, dass die Re-ID nur für einen kurzen Moment fehlerhaft ist, was bspw. durch Mittelung über mehrere Frames gefiltert werden kann. Im Vergleich zum OR1-Datensatz verbessern die farbigen OP-Hauben die Zuordnung, sodass eine zuverlässige Re-Identifikation mehrheitlich möglich ist.

Das Experiment hat aber auch gezeigt, dass durch Finetuning eines Modells die Robustheit gesteigert werden kann, vorausgesetzt eine ausreichend große Datenmenge zum Training ist vorhanden. Somit ist, wie auch in der gängigen Praxis üblich, Finetuning von Modellen nach Möglichkeit einem Training mit vermeintlich gut generalisierbaren, aber nicht domänenspezifischen Daten zu bevorzugen.

Da die Datensätze OR0, OR1 und OR2 alle die gleiche Personengruppe zeigen, alle Personen die gleichen Aufgaben ausführen und alle Datensätze in der gleichen Umgebung aufgenommen wurden, kann davon ausgegangen werden, dass die Unterschiede in den Testergebnissen tatsächlich auf die unterschiedlichen Kleidungsstile in den einzelnen Datensätzen zurückzuführen sind. Aufgrund

der zu erwartenden Ergebnisse und der vorhandenen Argumente bzgl. der Abweichungen, ist außerdem anzunehmen, dass der Datensatz geeignet ist, um einen ersten Überblick über die Auswirkungen von medizinischer Kleidung zu schaffen. Für tiefergehende Studien und allgemeingültige Aussagen sind jedoch größere Datensätze erforderlich.

Entsprechend kann zusammengefasst werden, dass einheitliche Kleidung und insbesondere medizinische Kleidung starke Auswirkungen auf die Leistung von Re-ID-Algorithmen hat, die nicht speziell mit Bildern aus diesem Bereich trainiert wurden. Medizinische Kleidung bedeckt große Teile des Körpers von Personen und damit Unterscheidungsmerkmale, die für deren Re-Identifikation wichtig sind. Zusätzliche Markierungen in Form von farbigen OP-Hauben können die Ergebnisse verbessern, gehen aber nicht über die Referenz mit Alltagskleidung hinaus, weshalb nach wie vor ein Finetuning der einzusetzenden Modelle empfohlen wird.

5.4 Diskussion & Fazit zum Teilsystem zur Aktivitätsanalyse innerhalb des OP-Saals

In diesem Kapitel wurde ein System zur Analyse menschlicher Bewegungen in Videoaufnahmen beschrieben. Grundlage für die Analyse bildet die Berechnung sog. Bewegungsvektoren, welche sich mittels Posenerkennung und -tracking aus der Beobachtung von Gelenkpunkten einer Person über die Zeit ergeben (Anforderungen [*AF-04*]). Darauf basierend wurde eine mathematische Methode erarbeitet, mit der verschiedene Bewegungsarten detektiert werden können. Die Unterscheidung der Bewegungsarten erfolgt dabei maßgeblich über unterschiedliche Auswertezeiträume der Bewegungsvektoren. Außerdem können mit diesem Ansatz sowohl die Bewegungen einzelner Personen als auch die Gesamtaktivität (also die Bewegungen aller Personen) im Raum ausgewertet werden, was im Gesamtsystem bei der Erkennung von Anomalien oder Verzögerungen im Ablauf helfen kann (Anforderung [*AF-03*]).

Ein Problem bei der Umsetzung des Konzepts war die eindeutige Zuordnung von Gelenkpunkten zu den Personen über mehrere Kameraframes hinweg. Dies führte zunächst zu fehlerhaften Berechnungen aufgrund vermischter Bewegungen. Durch Einführung eines Trackingmechanismus für die erkannten Posen konnte dieses Problem jedoch gelöst werden (Anforderung [*AF-05*]).

Eine besondere Herausforderung beim Tracking von Personen im medizinischen Umfeld stellt die Verdeckung vieler Erkennungsmerkmale durch medizinische Kleidung dar. Aus diesem Grund wurde in diesem Kapitel zusätzlich der Einfluss auf bestehende Re-ID-Methoden untersucht und inwiefern eindeutige Marker an Personen zur Problemlösung beitragen können. Dazu wurden fünf verschiedene Re-ID-Modelle aus dem Stand der Technik, die nur mit Probanden in Alltagskleidung trainiert wurden, mit eigens aufgezeichneten Daten aus einem simulierten medizinischen Umfeld getestet. Das Resultat zeigt, dass die Diskriminierungsgenauigkeit dabei ohne weiteres Training oder Markierungen um mindestens 13 Prozentpunkte in der Rank-1-Genauigkeit bzw. 30 Prozentpunkte in der mAP abfällt. Mit eindeutigen farblichen Markierungen, aber weiterhin ohne zusätzliches Training, fällt die Rank-1-Genauigkeit im besten Fall lediglich um 4,3 Prozentpunkte und die mAP um 62,9 Prozentpunkte. Es konnte außerdem gezeigt werden, dass Finetuning der Modelle mit dem realen Anwendungsfeld ähnlicheren Daten zu weiteren Verbesserungen führt.

Aufgrund fehlender Daten aus echten OP-Sälen konnten die Konzepte nicht vollumfänglich umgesetzt und evaluiert werden. Mit den entsprechenden Daten müssen in zukünftigen Arbeiten die Auswertemethoden weiter spezifiziert und ggf. zusätzliche Filter eingesetzt werden, um die Bewegungen möglichst genau zu erfassen. Damit kann dann auch die Hypothese, dass eine gesteigerte Gesamtaktivität auf Unregelmäßigkeiten und somit auch auf Verzögerungen im Ablauf hindeutet, konkret beantwortet werden. Darüber hinaus bietet ein robustes System zur Beobachtung von OP-Sälen weitere Unterstützungsmöglichkeiten für das OP-Team. Bspw. kann ein solches System auch bei der Arbeitssicherung helfen, indem die Gesamtzahl anwesender Personen beim Röntgen oder der Abstand der Anwesenden zum Röntgengerät detektiert wird und entsprechende

Warnungen ausgegeben werden. Eine weitere mögliche Erweiterung ist die Betrachtung ergonomischer Aspekte, wie Achsveränderungen in der Körperhaltung, was generelle Rückschlüsse zur Verbesserung des Raumdesigns zulässt.

6 Teilsystem: Detektion von OP-Materialien am Instrumententisch

Kapitel 6 stellt die Arbeiten zum Teilsystem am Instrumententisch näher dar. Hierfür wird zunächst das Konzept aus Kapitel 4 genauer ausgearbeitet. Im Anschluss wird die Erstellung eines geeigneten Datensets unter Laborbedingungen ausgearbeitet. Darauf aufbauend werden verschiedene Modellarchitekturen zur Objektdetektion aus dem Stand der Technik zum Training mit den für den vorliegenden Anwendungsfall geeigneten Daten ausgewählt und trainiert. Diese werden abschließend bzgl. ihrer Nutzbarkeit im zugrundeliegenden Gesamtkonzept evaluiert.

6.1 Systemkonzept zur Materialdetektion am Instrumententisch

Wie in den vorherigen Abschnitten zur Prozessanalyse und zum Gesamtkonzept bereits erläutert, ist die Grundidee des Teilsystems zur Erkennung der Instrumente am Instrumententisch die Identifikation der vergangenen und der aktuell laufenden OP-Phasen auf Basis der zeitlichen Abfolge der genutzten Instrumente und Materialien. Wesentliche Voraussetzung dafür ist deren robuste Erkennung im Kamerabild. Dafür muss der technische Aufbau des Systems so umgesetzt sein, dass der gesamte Instrumententisch kontinuierlich erfasst werden kann und Verdeckungen, bspw. durch das OP-Personal, weitestgehend verhindert werden.

Eine Möglichkeit dafür ist die Anbringung der Kamera direkt am Instrumententisch. Nachteil dabei ist allerdings, dass einerseits die Stromversorgung und Anbindung der Kamera ans Krankenhausnetzwerk erschwert wird und andererseits die hygienischen Anforderungen (z. B. Desinfizierbarkeit der Kamera) hoch sind. Darüber hinaus erschwert die Anbringung am Tisch die Erreichbarkeit der darauf liegenden Objekte von allen Seiten aus. Eine alternative Möglichkeit der Anbringung ist an der Decke des Raumes, wodurch alle genannten Probleme gelöst werden. Gleichzeitig steigt jedoch das Risiko der Verdeckung durch die größere Distanz zwischen Kamera und Instrumententisch. Darüber hinaus muss die Kameratechnik höheren Anforderungen genügen, da die größere Entfernung eine höhere Auflösung oder Zoom-Möglichkeiten erfordert. Zudem erschwert die feste Kameraposition die Erfassung des beweglichen Instrumententischs, da dieser nicht immer exakt an der gleichen Stelle steht oder während der OP bewegt werden kann. Ein ausreichend großer Aufnahmewinkel oder die Möglichkeit der Motorsteuerung zur Kameraausrichtung können dieses Problem lösen. Infolge der geringeren Einschränkungen für das OP-Personal und Auswirkungen auf die bestehenden Abläufe wird in der weiteren Konzeptionierung der Ansatz der Deckenkamera verfolgt.

Algorithmisch soll die Instrumentennutzung am Instrumententisch dadurch detektiert werden, dass das System erfasst, wenn das entsprechende Instrument nicht mehr auf dem Tisch liegt, also aus dem Kamerabild verschwindet. Als logische Konsequenz daraus ergibt sich, dass es durch das OP-Team in Benutzung ist. Nach der Nutzung wird es an seinen ursprünglichen Platz zurückgelegt und kann wieder als „nicht in Benutzung“ markiert werden. Das Flussdiagramm in Abbildung 6.1 skizziert diese Zusammenhänge. Für die Erkennung der Instrumente soll ein Deep-Learning-basierter Ansatz zur Objekterkennung nach dem aktuellen Stand der Technik genutzt werden, welcher auf die zu erkennenden Instrumente trainiert wird. Relevant für die Auswahl eines geeigneten Modells sind, neben seiner Erkennungsgenauigkeit und -geschwindigkeit bei der Inferenz, auch die Fähigkeit alle Objekte auf dem Instrumententisch gleichzeitig zu erkennen. Außerdem müssen aufgrund fehlender öffentlicher Daten eigene Trainingsdaten erzeugt werden. Um den Aufwand hierfür möglichst gering zu

halten, sollte das verwendete Modell in der Lage sein, mit einer geringen Anzahl an Trainingsdaten gute Ergebnisse zu erreichen. Die Auswahl des Modells wird in Abschnitt 6.3 genauer beschrieben.

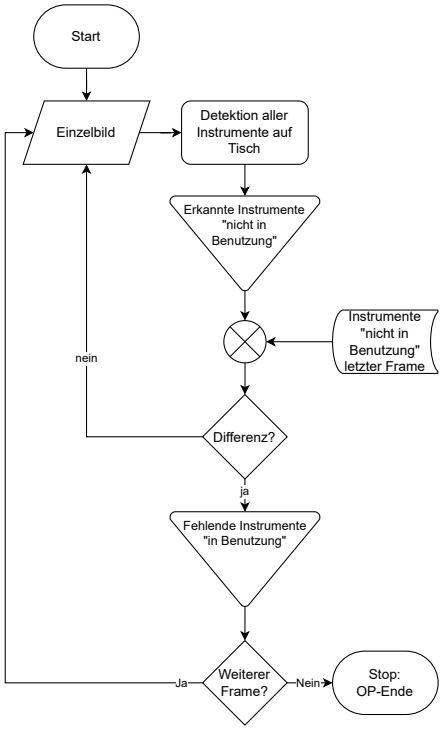


Abbildung 6.1: Flussdiagramm zur Erkennung der Instrumentennutzung am Instrumententisch

6.2 Laborsetup & Datenakquise

Zunächst muss ein entsprechender Datensatz aufgebaut werden, da keine Daten zum Training eines Erkennungssystems für laparoskopische OP-Instrumente aus der Perspektive von Deckenkameras öffentlich verfügbar sind. Datensets wie

Cholec80 oder HeiChole (vgl. Abschnitt 3.4.2) beinhalten zwar grundsätzlich annotierte Videos mit Instrumenten, allerdings sind diese aus Perspektive der Endoskopkamera aufgezeichnet, sodass sie für den genannten Anwendungsfall nicht genutzt werden können. Im angedachten System soll der gesamte Instrumententisch im Bild sichtbar sein. Die einzelnen Instrumente sind also in kompletter Ansicht enthalten, dafür aber wesentlich kleiner als aus der Endoskop-Perspektive. Dies führt dazu, dass die Maulteile, die oftmals das Hauptunterscheidungsmerkmal darstellen, nur schwer unterscheidbar sind. Aus diesem Grund muss auch eine zu den genannten öffentlichen Datensets unterschiedliche Klassifizierung der einzelnen Instrumente vorgenommen werden.

Für die eigene Datenaufzeichnung stehen insgesamt 36 verschiedene Instrumente für die laparoskopische Cholezystektomie der Richard Wolf GmbH[1] zur Verfügung. Diese lassen sich gemäß ihrer Optik und Funktion in insgesamt 15 Klassen kategorisieren. Diese Klassen sowie die entsprechende Klasse im HeiChole-Datensatz und die OP-Phasen, in denen das Instrument üblicherweise eingesetzt wird, sind in Tabelle A.6 dargestellt. Um repräsentative Daten aufzeichnen zu können, ohne dass der Aufwand für eine Integration eines Systems in eine reale OP-Umgebung betrieben werden muss (z. B. bzgl. Anforderungen zur Regulatorik oder Hygiene), wurde zunächst ein Laborsystem aufgebaut, welches eine solche Umgebung nachbildet. Dies ist im Rahmen der Machbarkeitsstudie für das Erkennungssystem ausreichend. Für die Umsetzung und Evaluation des Gesamtsystems ist jedoch ein umfangreicherer Datensatz aus realen OP-Umgebungen notwendig. Der Laboraufbau wurde durch Montage der im Gesamtkonzept bereits genannte Canon VB-H45 Netzwerk-Kamera über einem Tisch umgesetzt. Anschließend wurde die Kamera mithilfe der Motorsteuerung und dem optischen Zoom so ausgerichtet, dass der komplette Tisch aufgezeichnet wird und möglichst viel Platz im Kamerabild einnimmt. Der Tisch wurde mit einem einfarbigen Tuch als Ersatz der sterilen Abdeckung realer OPs abgedeckt, auf dem dann die Instrumente einzeln oder in Gruppen angeordnet

[1] `https://www.richard-wolf.com/`

für die Aufnahmen platziert wurden. Der gesamte Aufbau ist in Abbildung 6.2 dargestellt.

Abbildung 6.2: Laboraufbau des Instrumententisch-Systems

Für die Datenaufzeichnung wurden zur Erhöhung der Diversität der Daten verschiedene Ausrichtungen der Instrumente und Hintergründe vorgesehen. Bei der Ausrichtung wurde nicht nur die Orientierung auf dem Tisch, sondern auch Besonderheiten der Instrumente, wie bspw. der Öffnungszustand von Zangen, und Verdeckung durch Überlappung berücksichtigt (vgl. Abbildung 6.3). Auf diese Weise wurden insgesamt 21.913 Einzelbilder erstellt und annotiert.

Das gesamte Datenset wurde anschließend zu ca. 60 Prozent für das Training, 20 Prozent für das Testen und 20 Prozent für die Validierung der Modelle aufgeteilt. Aus Abbildung 6.4 wird ersichtlich, dass die einzelnen Klassen dabei nicht

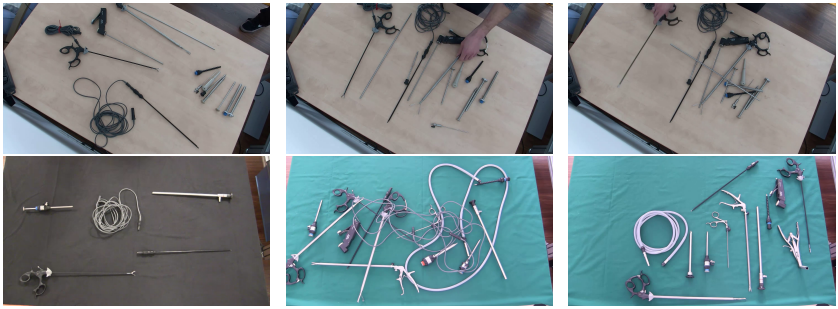


Abbildung 6.3: Beispielbilder des aufgezeichneten Datensets zur Instrumentenerkennung

gleichmäßig verteilt sind. Dies liegt hauptsächlich daran, dass im zugrundeliegenden Instrumentenset für einige Klassen unterschiedliche Ausprägungen vorhanden sind (bspw. Optik 10 mm 0°, Optik 10 mm 30° und Optik 5,2 mm 30°, vgl. Tabelle A.6) und durch gleichzeitiges Vorhandensein mehrerer Varianten während des Aufnahmeprozesses mehr Klasseninstanzen pro Einzelaufnahme produziert wurden. Dies führt gleichzeitig auch dazu, dass die Verteilung eher der Realität im OP-Saal entspricht als eine gleichmäßige Verteilung über alle Klassen hinweg. Insofern sollte diese Tatsache sogar förderlich für den realen Einsatz eines damit trainierten Modells sein. Eine detaillierte Darstellung der Verteilung der Klassen und der Aufteilung in Trainings-, Test- und Validierungsdaten ist Tabelle A.7 zu entnehmen.

6.3 Modellauswahl & Trainingsprozess

Für die Evaluation geeigneter Objekterkennungsmodelle zur Auswertung am Instrumententisch wurden im Wesentlichen verschiedene Variationen der YOLO-Familie ausgewählt. Diese bieten üblicherweise einen guten Kompromiss zwischen Erkennungsqualität und Geschwindigkeit und erfüllen alle zuvor diskutierten Anforderungen. Ferner sind durch die verschiedenen Varianten über die Zeit unterschiedliche Architekturansätze in die Entwicklung eingeflossen, deren Eignung im vorliegenden Anwendungsfall untersucht werden soll. Ergänzt

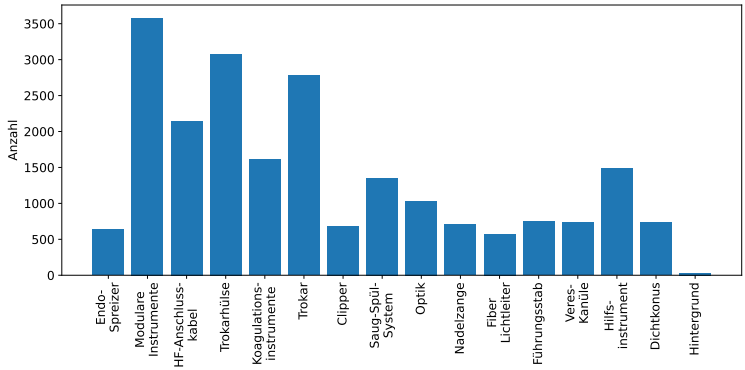


Abbildung 6.4: Verteilung der Instrumenten-Klassen im Datenset

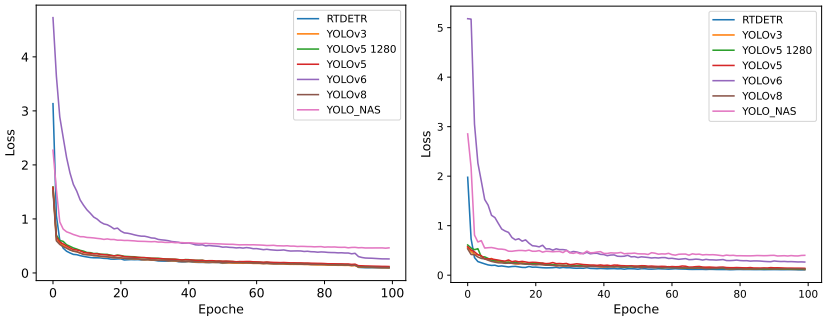
wurde die Auswahl durch ein transformerbasiertes Modell, da diese Architektur in der aktuellen Forschungslandschaft viel diskutiert ist und schnelle Fortschritte erreicht. Alle Modelle wurden in unterschiedlichen Skalierungsvarianten untersucht. Die vollständige Liste ist Tabelle 6.2 zu entnehmen.

Für eine bessere Vergleichbarkeit wurden alle Modelle auf dem gleichen System (vgl. System 2 in Tabelle A.1) und möglichst mit den Standard-Hyperparametern des Trainings-Frameworks trainiert. Da kein einzelnes Framework für alle zu trainierenden Modelle verfügbar war, wurden sowohl *Ultralytics* [85] als auch *SuperGradients* [3] verwendet. Letzteres wurde dabei lediglich für die Varianten von YOLO-NAS eingesetzt. Alle anderen Modelle konnten mit Ultralytics trainiert werden. Die wichtigsten Parameter sind in Tabelle 6.1 aufgelistet. Die Festlegung orientierte sich stark an den Standardwerten des jeweiligen eingesetzten Frameworks. Eine Ausnahme zu den in der Tabelle genannten Werten bilden die Modelle zu YOLOv5. Hier wurden alle Skalierungsvarianten zusätzlich mit einer Bildgröße von 1280 x 1280 Pixeln trainiert. Dadurch kann untersucht werden, ob der höhere Detailgrad der Bilder zu Verbesserungen in der Erkennungsqualität führt und wie sich das auf die Performance des Modells auswirkt.

Tabelle 6.1: Übersicht verschiedener Parameter für das Training und die Validierung der Modelle zur Instrumentendetektion am OP-Tisch.

	Parameter	Ultralytics	SuperGradients
Data	Image-Size	640	640
	Augmentierung	false	Mosaic, Random Affine, Mixup, HSV, Horizontal Flip, Padded Rescale
Train	Optimizer	AdamW	Adam
	Learning Rate	0,000526	0,0005
	Weight Decay	0,0005	0,0001
	Epochen	100	100
	Batch-Size	16	32
Test	Konfidenz-Schwellwert	0,001	0,01
	NMS-Schwellwert	0,6	0,7

Abbildung 6.5 stellt den Verlauf der Trainings- und Validierungsverluste dar. Zur besseren Übersichtlichkeit wurde pro Modell beispielhaft lediglich eine Variante geplottet. Die Verläufe deuten auf einen korrekten Trainingsablauf hin und lassen kein Overfitting vermuten.



(a) Training **(b)** Validierung

Abbildung 6.5: Trainingsverlauf der trainierten Modelle am Instrumententisch

6.4 Evaluation der trainierten Modelle zur Instrumentendetektion

Die Evaluationsergebnisse für die einzelnen Modelle sind in Tabelle 6.2 dargestellt. Dabei ist zunächst festzuhalten, dass aufgrund der unterschiedlich verwendeten Trainingsframeworks teilweise leicht unterschiedliche Informationen zur Verfügung stehen. So konnten zu den Modellvarianten zu RT-DETR keine Angaben zum Berechnungsaufwand extrahiert werden. Außerdem wurde dieser bei den YOLO-NAS Modellen nicht in „Giga Floating Point Operations (GFLOPs)“ sondern in „Giga Multiply–Accumulate Operations (GMACs)“ berechnet. Da sich 1 MAC aus der Berechnungsformel $a \cdot x + b$ ergibt, ist der Zusammenhang näherungsweise 1 GMACs $\approx 0,5 \cdot$ GFLOPs, so dass eine Vergleichbarkeit hergestellt werden kann. Die Originalwerte in GMACs sind jeweils in Klammern mit in der Tabelle angegeben.

Bei der Betrachtung der Modellstrukturen wird ersichtlich, dass diese sich mit ca. 2,5 Mio. bis ca. 111 Mio. Parametern und einem Berechnungsaufwand zwischen ca. 7 und ca. 392 GFLOPs stark unterscheiden. Das wiederum hat Einfluss auf die jeweilige Inferenzzeit, welche auf dem Testsystem zwischen 0,605 ms und 8,520 ms lag. Weiterhin fällt auf, dass die Werte für mAP nur kleine Unterschiede aufweisen. So ist die maximale Differenz lediglich 0,0087 Punkte bei der mAP bei fester IoU von 0,5 bzw. 0,0687 Punkte bei der mAP im Bereich [0,5:0,95]. Ein ähnliches Bild zeigt auch der F1-Score. Hier fallen allerdings die YOLO-NAS-Modelle mit einer Abweichung von bis zu 0,538 Punkten stark ab. Die restlichen Modelle unterscheiden sich vom besten Modell lediglich um maximal 0,0217 Punkte. Der Grund für die große Abweichung der YOLO-NAS-Varianten liegt in deren schlechten Präzision, die stets kleiner als 0,5 ist. Dies zeigt, dass YOLO-NAS viele falsch-positive Vorhersagen produziert. Auch die YOLOv6-Varianten schneiden, wenn auch weniger ausgeprägt, deutlich schlechter ab, als YOLOv3, YOLOv5, YOLOv8 und RT-DETR. Hier sind im Gegensatz zu YOLO-NAS allerdings zahlreiche falsch-negative Vorhersagen, also Instrumente, die der falschen Klasse zugeordnet wurden, und somit ein schlechterer Recall ursächlich. Die transformerbasierten RT-DETR-Modelle

Tabelle 6.2: Übersicht über alle trainierten Modelle zur Objekterkennung am Instrumententisch

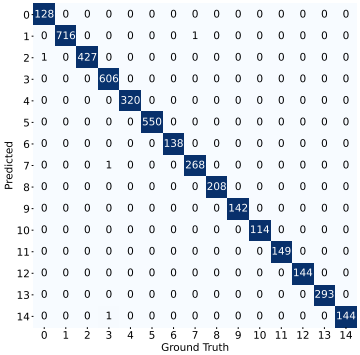
Modell	mAP @0,5	mAP @[0,5:0,95]	Precision	Recall	F1-Score	Berechnungsaufwand (GFLOPs)	Parameter	Inferenzzeit (ms)
YOLOv3-tinyu	0,9917	0,9682	0,9891	0,9850	0,9870	19,1	12,1 M	**0,605**
YOLOv3-u.pt	**0,9948**	0,9869	0,9934	0,9921	0,9928	283,0	103,7 M	3,488
YOLOv3-sppu	**0,9948**	0,9869	0,9957	0,9935	0,9946	283,9	104,8 M	3,561
YOLOv5-nu	0,9940	0,9674	0,9892	0,9878	0,9885	**7,2**	**2,5 M**	0,751
YOLOv5-su	0,9947	0,9831	0,9936	0,9953	0,9944	24,1	9,1 M	1,066
YOLOv5-lu	**0,9948**	0,9872	0,9921	**0,9961**	0,9941	135,3	53,2 M	2,359
YOLOv5-xu	**0,9948**	0,9879	0,9950	0,9951	0,9950	247,0	97,2 M	3,701
YOLOv5-n6u 1280	0,9947	0,9815	**0,9971**	0,9904	0,9937	7,3	4,1 M	1,792
YOLOv5-s6u 1280	0,9947	0,9870	0,9952	0,9957	**0,9955**	24,5	15,3 M	2,970
YOLOv5-m6u 1280	0,9947	0,9879	0,9950	0,9950	0,9950	65,5	41,2 M	5,328
YOLOv5-l6u 1280	0,9947	0,9878	0,9928	0,9946	0,9937	137,7	86,0 M	8,520
YOLOv5-x6u 1280	Trainingsabbruch aufgrund von zu geringem Grafikspeicher.							
YOLOv6n	0,9861	0,9194	0,9840	0,9639	0,9738	11,9	4,2 M	0,838
YOLOv6s	0,9936	0,9600	0,9916	0,9773	0,9844	44,2	16,3 M	1,074
YOLOv6m	0,9931	0,9656	0,9895	0,9789	0,9842	161,6	52,0 M	2,088
YOLOv6l	0,9909	0,9644	0,9898	0,9750	0,9824	392,0	110,9 M	3,681
YOLOv8n	0,9944	0,9740	0,9924	0,9891	0,9908	8,2	3,0 M	0,823
YOLOv8s	**0,9948**	0,9853	0,9931	0,9937	0,9934	28,7	11,1 M	1,059
YOLOv8m	**0,9948**	0,9873	0,9954	0,9954	0,9954	79,1	25,9 M	1,793
YOLOv8l	**0,9948**	**0,9881**	0,9957	0,9952	0,9954	165,5	43,6 M	2,610
YOLOv8x	**0,9948**	0,9879	0,9941	0,9954	0,9948	258,2	68,2 M	3,732
YOLO-NAS-s	0,9882	0,9353	0,2970	0,9957	0,4575	32,7 (16,36)	12,2 M	-
YOLO-NAS-m	0,9901	0,9528	0,4845	0,9943	0,6515	88,7 (44,37)	31,9 M	-
YOLO-NAS-l	0,9902	0,9589	0,4790	0,9941	0,6465	120,8 (60,41)	42,0 M	-
RT-DETR-l	0,9942	0,9812	0,9948	0,9933	0,9941	-	32,8 M	4,155
RT-DETR-x	0,9947	0,9828	0,9942	0,9941	0,9941	-	67,3 M	6,302

sind zwar von den Erkennungsraten nahe an den besten Ergebnissen, sind dafür aber größer und entsprechend auch erheblich langsamer. So benötigt YOLOv5-s6u 1280, welches den höchsten F1-Score aufweist, für die Inferenz 2,97 ms, wohingegen RT-DETR-x mit 6,302 ms mehr als doppelt so lange dauert. Die Modellgröße unterscheidet sich dabei nahezu um den Faktor 4,5.

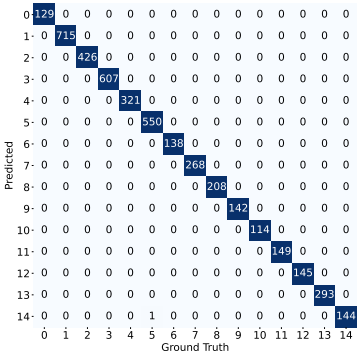
Die Untersuchung der unterschiedlichen Skalierungen der Modelle zeigt wie erwartet, dass die kleinen tiny- bzw. nano-Varianten allesamt mit einer Inferenzzeit von deutlich unter 1 ms sehr schnell sind. Dieser Vorteil geht allerdings mit signifikant schlechterer Erkennungsqualität einher. Durchschnittlich fällt der F1-Score um 0,62 Prozent und die mAP um 2,17 Prozent im Vergleich zum jeweils besten Modell ab. Umgekehrt sind allerdings nicht immer die größten Modellvarianten die genauesten. Dies zeigt sich sowohl bei YOLOv5 mit einer Bildgröße von 1280 x 1280 Pixeln wie auch bei YOLOv6, YOLOv8 und YOLO-NAS. Bezogen auf die Inferenzzeit sind wiederum die größten Varianten auch jeweils die langsamsten. Folglich bieten die Modellvarianten s, m und l den besten Kompromiss aus Geschwindigkeit und Genauigkeit. Wie bereits erwähnt, erreichte YOLOv5-s6u 1280 von allen getesteten Modellen den höchsten F1-Score und befindet sich auch bei den übrigen Metriken im oberen Bereich. Die Inferenzzeit ist mit fast 3 ms im mittleren Bereich des Tests. Allerdings bringt die größere Bildgröße signifikant höhere Anforderungen an die Hardware mit sich, da einerseits für den Datenaustausch höhere Bandbreiten und andererseits für die Datenhaltung mehr Speicher erforderlich sind. Die Herausforderung wird auch dadurch deutlich, dass das gleiche Modell in der größten Variante aufgrund von zu geringem GPU-Speicher im Testsystem nicht zu Ende trainiert werden konnte. In einem ähnlichen Wertebereich bei allen berechneten Ergebnissen befinden sich die Modelle YOLOv8m und YOLOv8l. Hier fallen der F1-Score und der Recall minimal schlechter aus, dafür sind die mAP sowie die Precision etwas höher. Außerdem ist die Inferenzzeit insb. bei der m-Variante geringer.

Abbildung 6.6 zeigt die Konfusionsmatrizen der drei Modelle, die die besten Evaluationsergebnisse erzielen konnten. Für die restlichen Modelle sind die jeweiligen Grafiken im Anhang A.3.2 zu finden. Insgesamt sind hier nur äußerst geringe Unterschiede festzustellen, sodass keine Rückschlüsse darüber getroffen werden können, welches Modell möglicherweise am besten für einen realen Einsatz geeignet wäre. YOLOv8m zeigt zwar die wenigsten Fehldetektionen, dies aber in einem vernachlässigbaren Bereich.

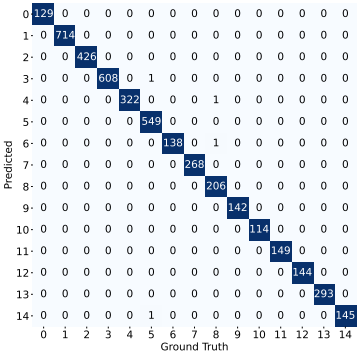
Die Graphen in Abbildung 6.7 skizzieren für die gleichen Modelle die Precision-Recall-Kurven, um das Modell mit dem geringsten Kompromiss zwischen



(a) YOLOv5s6u



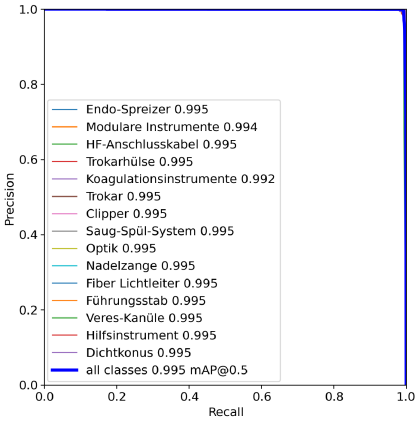
(b) YOLOv8m



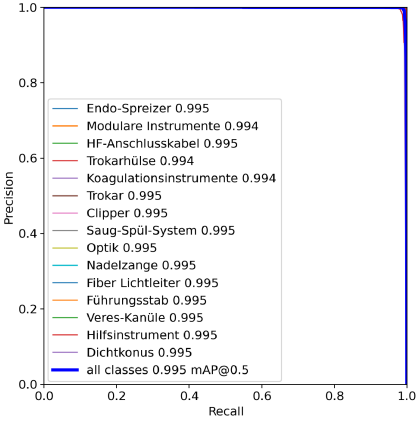
(c) YOLOv8l

Abbildung 6.6: Konfusionsmatrizen zu den Modellen YOLOv5-s6u 1280, YOLOv8m und YOLOv8l. Die Labels entsprechen der Nummerierung in Tabelle A.7 im Anhang.

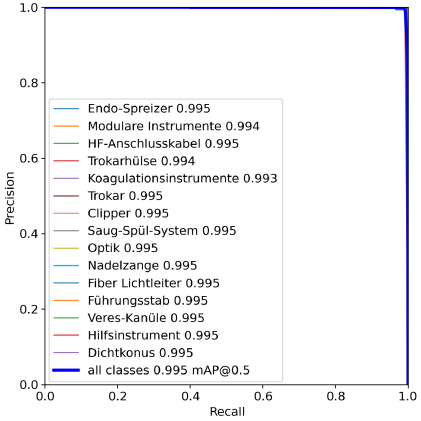
Precision und Recall zu finden. Auch hier sind keine signifikanten Unterschiede auszumachen, die für oder gegen eines der Modelle sprechen.



(a) YOLOv5s6u



(b) YOLOv8m



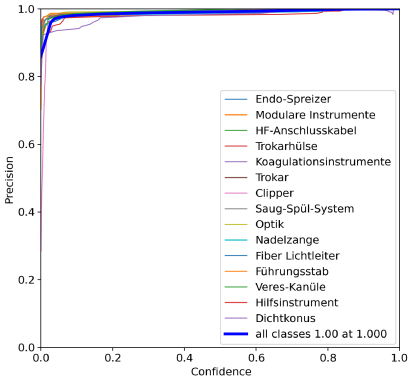
(c) YOLOv8l

Abbildung 6.7: Precision-Recall-Kurven zu den Modellen YOLOv5-s6u 1280, YOLOv8m und YOLOv8l

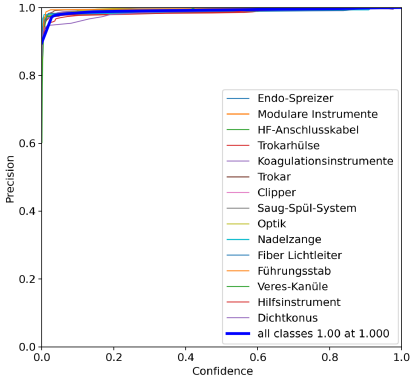
Auch tiefergehende Untersuchungen mit Hilfe der Abbildungen 6.8 bis 6.10 der Precision-, Recall- und F1-Kurven, die die jeweilige Metrik in Abhängigkeit der Konfidenz darstellen, erlauben keine eindeutige Festlegung des am besten geeigneten Modells. Die P-Kurven zeigen für YOLOv8m (Abb. 6.8b) die kleinste Streuung über alle Klassen hinweg und den schnellsten Anstieg nahe des Maximums bei einer Konfidenz von ungefähr 0,6. YOLOv8l (Abb. 6.8c) erreicht das erst bei einer Konfidenz von 0,9, YOLOv5-s6u (Abb. 6.8a) zwar bereits bei einer Konfidenz von 0,8, hat dafür aber nochmal einen Ausreißer ab ca. 0,95. Bei den R-Kurven wird ersichtlich, dass beide Varianten von YOLOv8 über alle Klassen hinweg ab einer Konfidenz von ungefähr 0,8 schnell abfallen. Dabei zeigt YOLOv8m (Abb. 6.9b) insgesamt die breiteste Streuung vor der abfallenden Phase der drei Modelle. Merklich fallen einzelne Klassen ab einer Konfidenz von 0,58. Dies tritt bei YOLOv8l (Abb. 6.9c) bereits ab einer Konfidenz von ca. 0,42 und bei YOLOv5-s6u (Abb. 6.9a) sogar schon ab ca. 0,3 auf. Auch gemittelt über alle Klassen sinkt der Recall bei letztgenanntem Modell am frühesten. Die F1-Kurven stützen die beschriebenen Beobachtungen. Die Kurven sehen für alle drei Modelle sehr ähnlich aus. YOLOv5-s6u erreicht dabei den niedrigsten rechnerisch optimalen Konfidenz-Grenzwert über alle Klassen hinweg von 0,693 und zeigt auch die größten Ausreißer bei den einzelnen Klassen (vgl. Abb. 6.10a). Die beiden YOLOv8-Varianten liegen mit Grenzwerten von 0,721 für das m-Modell und 0,745 (Abb. 6.10b) für das l-Modell (Abb. 6.10c) etwas höher als das YOLOv5-Modell. Die l-Variante zeigt dabei die besseren Werte bei kleinen Konfidenzen von weniger als 0,1, die m-Variante scheint dafür auch bei Konfidenzen oberhalb des optimal berechneten Wertes bis zu einer Konfidenz von ca. 0,9 noch stabiler zu sein.

6.5 Diskussion & Fazit zum Teilsystem zur Instrumentendetektion

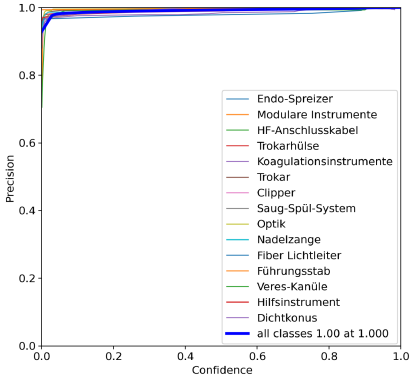
Die vorangegangenen Abschnitte legten dar, dass die getesteten Modelle allesamt eine gute Erkennungsqualität und die Ergebnisse überwiegend nur geringe



(a) YOLOv5s6u



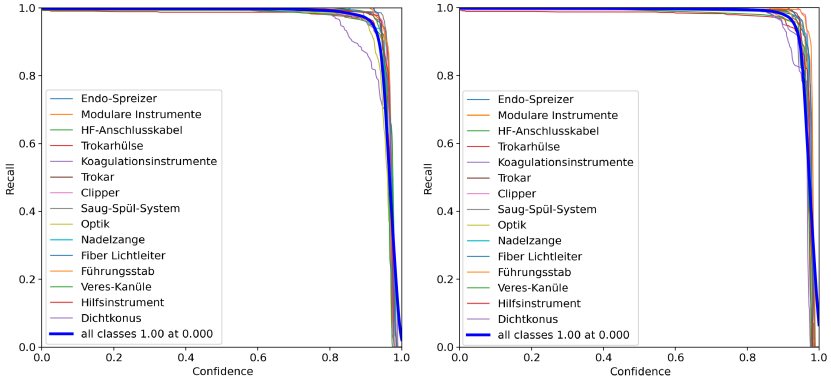
(b) YOLOv8m



(c) YOLOv8l

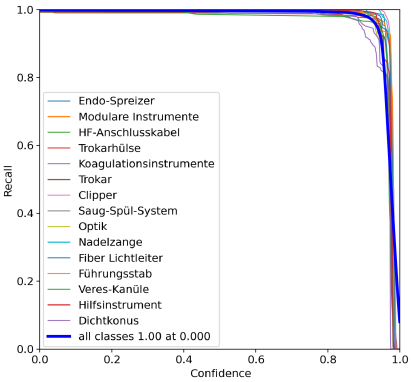
Abbildung 6.8: Precision-Kurven zu den Modellen YOLOv5-s6u 1280, YOLOv8m und YOLOv8l

Unterschiede aufweisen. Lediglich YOLO-NAS scheint signifikant schlechter geeignet, was insb. auf die geringere Precision zurückzuführen ist. Diese Tatsache macht YOLO-NAS im Gesamtsystem nicht einsetzbar, da die vielen falsch-positiven Datenpunkte großen Einfluss auf die zur Instrumentenerkennung nachgelagerten Berechnungen hätte und sich der jeweilige Fehler somit immer weiter fortpflanzen würde.



(a) YOLOv5s6u

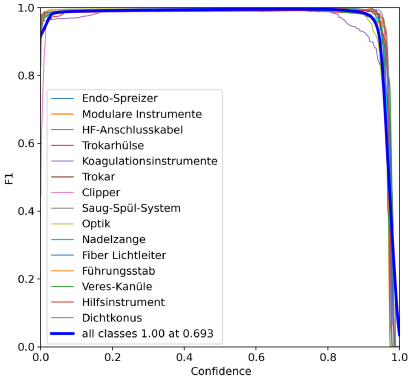
(b) YOLOv8m



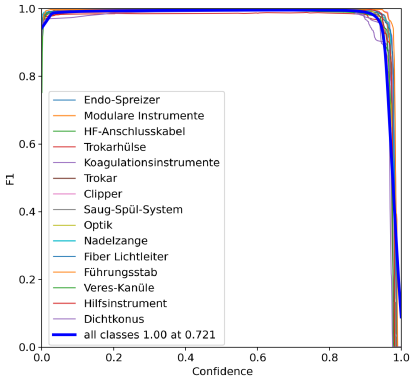
(c) YOLOv8l

Abbildung 6.9: Recall-Kurven zu den Modellen YOLOv5-s6u 1280, YOLOv8m und YOLOv8l

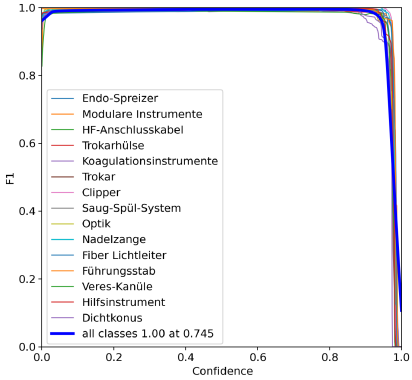
RT-DETR erreicht vergleichbare Erkennungsergebnisse, ist dabei allerdings langsamer als die besten evaluierten Modelle und insgesamt auch größer, was höhere Hardwareanforderungen mit sich bringt. Der moderne transformerbasierte Erkennungsansatz erlaubt die parallele Verarbeitung von mehreren Objekten und kann dadurch besser komplexe Interaktionen zwischen Objekten berücksichtigen. Dies könnte im realen Einsatz vorteilhaft sein. In den hier durchgeführten Untersuchungen ergab sich allerdings kein merklicher Vorteil.



(a) YOLOv5s6u



(b) YOLOv8m



(c) YOLOv8l

Abbildung 6.10: F1-Kurven zu den Modellen YOLOv5-s6u 1280, YOLOv8m und YOLOv8l

Die genauere Untersuchung der drei besten Modelle zeigt nur minimale Unterschiede. Dabei erreicht YOLOv5-s6u 1280 zwar insgesamt den höchsten F1-Score, hat gleichzeitig aber auch bei einzelnen Klassen die größten Abweichungen sowohl in der P- als auch der R-Kurve generiert und auch den niedrigsten optimalen Konfidenz-Grenzwert erreicht. Dazu kommt, dass die größere Bildgröße erheblich höhere Anforderungen bzgl. Speicherbedarf und

Bandbreite verursacht. Die Systeme sind im eigenen Rechenzentrum der Krankenhäuser zwar prinzipiell gut skalierbar, dennoch spielen Kosten und auch Platzbedarf eine entscheidende Rolle bei der Umsetzung und Beschaffung neuer Gerätschaften. Somit wäre diese Tatsache für ein Realsystem definitiv nachteilig.

Zwischen den YOLOv8-Varianten scheint nach den durchgeführten Evaluationen die m-Variante insgesamt den besseren Kompromiss darzustellen. So ist das Modell kleiner, schneller und erreicht identische Werte beim F1-Score sowie der mAP@0,5. Zudem zeigt diese Version eine minimal bessere Präzision, was stabilere Ergebnisse in den nachgelagerten Berechnungen, wie der Phasenerkennung, verspricht. Dies erscheint zum Auswertezeitpunkt wichtiger für das Gesamtkonzept als der höhere Recall der l-Variante.

Als Fazit wird an dieser Stelle entsprechend YOLOv8 in der m-Variante als geeignetstes Modell für die Erkennung von OP-Instrumenten am Instrumententisch empfohlen (Anforderungen [*AF-06*] und [*AF-07*]). Da die Evaluation lediglich auf eigens dafür generierten Daten stattfand, sollte das Ergebnis allerdings mittels Realdaten aus dem Krankenhaus verifiziert werden. Möglicherweise treten hier einzelne Vor- oder Nachteile mehr in den Vordergrund, die auf die bessere Eignung anderer Modelle schließen lassen.

7 Teilsystem: OP-Phasenerkennung mittels Analyse endoskopischer Videos

Das Teilsystem zur Auswertung der Informationen aus dem Endoskop ermöglicht Analysen aus dem Körperinneren. Dies können einerseits einzelne Objekte, bspw. die OP-Instrumente aus anderer Perspektive, (Gewebe-)Strukturen, bspw. Fettgewebe oder die Gallenblase, oder konkrete Handlungsabläufe, bspw. das Fassen der Gallenblase mit dem Greifer, sein. Die genannten Informationen lassen dann wiederum auf den OP-Fortschritt schließen. Im nachfolgenden Kapitel wird zunächst der Schwerpunkt und das konkrete Konzept für die Endoskopauswertung in dieser Arbeit erläutert, bevor Spezifika bei dessen Umsetzung und die darauffolgende Evaluation der daraus entstandenen Modelle dargelegt und diskutiert werden.

7.1 Systemkonzept zur Phasenerkennung in endoskopischen Videos

Das Konzept, welches die vorliegende Arbeit verfolgt, sieht vor, dass zunächst Instrumente im Endoskopbild erkannt und anschließend über deren Nutzungsverlauf Rückschlüsse auf die OP-Phase gezogen werden. Vorteil dieses Ansatzes ist einerseits, dass für die Instrumentenerkennung Methoden der Objekterkennung eingesetzt werden können, welche in der Literatur und der aktuellen Forschung weit verbreitet sind und für die auch Trainingsdatensätze verfügbar sind, die

in vergleichbaren Arbeiten bereits genutzt wurden (vgl. Kapitel 3). Andererseits kann die Erkennung durch das Teilsystem zum Instrumententracking am Instrumententisch (siehe Kapitel 6) zusätzlich verifiziert und dadurch robuster gestaltet werden.

Der Schritt der Instrumentenerkennung liefert für die weitere Verarbeitung einen Vektor, der Informationen über das Auftreten der OP-Instrumente für die letzten 10 Bilder enthält. Für jedes Einzelbild wird das Auftreten der Instrumente durch einen multi-hot-codierten Vektor beschrieben (z. B. $[0, 0, 1, 0, 1, 1, 0]$, wobei eine 1 für das Auftreten eines bestimmten Instruments steht). Jeder Index im multi-hot-codierten Vektor repräsentiert einen Instrumententyp. Für jeden Vorhersagezyklus wird ein neuer Instrumentenvektor geladen, der als Eingabe für das Transformer-Netzwerk verwendet wird.

Die Inferenz auf die OP-Phase erfolgt anschließend durch ein Transformer-Netzwerk, welches die Phase für den aktuellen Frame bestimmt. Entgegen der Standard-Implementierung der Transformer-Architektur werden die konkatenierten detektierten Phasen nicht direkt als Eingabe in den Decoder-Block verwendet, sondern zunächst in ein Phasen-Histogramm übertragen, welches die bisher erkannten Frames für jede Phase akkumuliert und demnach deren Verteilung repräsentiert. Dieses Histogramm bildet dann den zusätzlichen Input in den Decoder für den folgenden Frame. Dadurch soll verhindert werden, dass das Netz fast immer die zuletzt erkannte Phase inferiert, was ein häufiges und gleichzeitig naheliegendes Verhalten darstellt. Bei einer durchschnittlichen Schnitt-Naht-Zeit von 59 Minuten nach [99] (vgl. Abschnitt 2.1.4) und einer Bildübertragungsrate von 30 Frames/Sekunde, besteht der Gesamtprozess aus 106.200 Einzelframes, wobei im Optimalfall lediglich sieben Phasenwechsel stattfinden. Dadurch ist die Wahrscheinlichkeit sehr hoch, dass im nächsten Frame die gleiche Phase zu sehen ist wie im aktuellen. Der zusätzliche Parameter bildet die Verteilung der Phasen über die Zeit mit in den Trainingsprozess ab und löst so dieses Problem. Darüber hinaus hat die Nutzung des Histogramms noch weitere Vorteile. Es enthält Informationen über alle bisherigen Vorhersagen und somit indirekt die Dauer der einzelnen Phasen, was zu einer generalisierteren Darstellung des Status der Operation führt. Weiterhin haben falsche Vorhersagen

des Transformers keine großen Auswirkungen auf die nächste Vorhersage, da die Verteilung des Histogramms durch einen falschen Eintrag weniger stark beeinflusst wird. Aufgrund der langsamen Änderung des „Gesamtbildes“ verfügt der Transformer über genügend Informationen, um sich selbst für die nächste Vorhersage zu korrigieren. Auf diese Weise wird ein Dominoeffekt, der zu falschen Vorhersagen führt, vermieden. Generell sind die Histogramme unabhängig von der Reihenfolge der Phasen. Es ist möglich, dass die Phasen in unterschiedlicher Reihenfolge und nicht rein sequenziell auftreten. Da die Reihenfolge der Phasen im Histogramm nicht sichtbar ist, können unerwartete Phasen entsprechend einfach zum Histogramm hinzugefügt werden.

Die Inferenz des Transformers setzt sich also aus dem Vektor aus der Instrumentenerkennung und dem Phasenhistogramm zusammen. Abbildung 7.1 stellt den Ablauf des beschriebenen Konzepts grafisch aufbereitet dar.

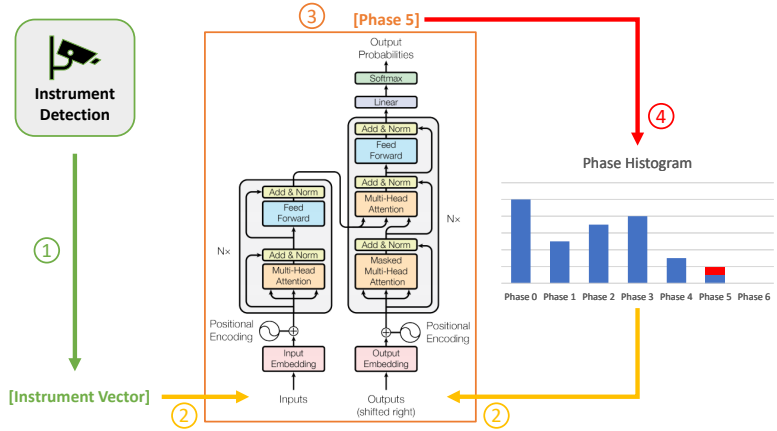


Abbildung 7.1: Darstellung des Konzepts zur OP-Phasenerkennung basierend auf der Erkennung der Instrumentennutzung & Transfomer-basierter zeitlichen Analyse:
Schritt 1: Instrumentenerkennung (grün), Schritt 2: Transformer-Eingabe Instrumente + aktuelles Histogramm (gelb), Schritt 3: Phasenprädiktion (orange), Schritt 4: Histogramm-Aktualisierung (rot).

7.2 Umsetzung der OP-Phasenerkennung

Wie im Konzept dargestellt, werden für die Umsetzung des Teilsystems zur Analyse der Endoskopvideos die Instrumentenerkennung und die darauf aufbauende Phasenerkennung zunächst getrennt voneinander betrachtet.

Für die Erprobung des hier diskutierten Konzepts wird eine robuste Intrumentenerkennung aus dem Endoskopbild vorausgesetzt. Da hierzu bereits ausgiebige Arbeiten aus unterschiedlichsten Forschungsgruppen verfügbar sind, soll in dieser Arbeit keine eigene Instrumentendetektion entwickelt werden. Stattdessen wurden vorhandene Ground Truth Daten aus den Cholec80- und HeiChole-Datensätzen (vgl. Abschnitte 3.4.2 und 3.4.2) an Stelle eines realen Objekterkenners eingesetzt. Dabei dienten die Annotationen zur Sichtbarkeit der Instrumente als Eingabedaten für das zu trainierende Transformer-Netzwerk und die Phasen-Annotationen als Ground Truth für die Netzwerkausgabe. Angesichts der Unterschiede in den beiden Datensets bzgl. Komplexität, Anzahl annotierter Einzelbilder und Gesamtanzahl betrachteter Eingriffe, wurde zunächst jeweils ein separates Modell mit jedem der beiden Datensets auf System 1 aus Tabelle A.1 im Anhang trainiert. Ziel dabei war es, die Auswirkungen der Unterschiede auf das finale Modell zu untersuchen. Erwartet wurden u. A. Differenzen in der Erkennungsqualität, aber auch der Generalisierbarkeit der Modelle.

Im Anschluss wurde ein Modell umgesetzt, welches die Daten aus beiden Datensätzen gleichzeitig zum Training nutzt. Aufgrund der gesteigerten Quantität und Varianz in den Daten war hierbei die Erwartungshaltung einer Verbesserung in der Erkennungsqualität und vor allem auch in der Generalisierbarkeit. Da die Annotationen der Instrumente nicht identisch sind, musste hierfür zunächst ein geeignetes Mapping gefunden und umgesetzt werden. Tabelle 7.1 zeigt die erarbeitete Zuordnung. Neben der uneinheitlichen Benennung des Saug-Spül-Rohrs, liegen die Unterschiede zum einen im Fehlen des „Staplers“ bei Cholec80 und zum anderen darin, dass sowohl der eingesetzte Haken als auch das Bipolar-Instrument zur Koagulation genutzt werden, was in HeiChole zusammengefasst als „Koagulationsinstrumente“ bezeichnet wird. Da der Stapler ohnehin nur in

einem einzigen Video auftritt, ist dessen Fehlen wenig relevant und kann in den Cholec80-Daten einfach immer als „nicht sichtbar“ gelabelt werden. Somit werden für das gemeinsame Training die Annotationen aus Cholec80 denen aus HeiChole angeglichen.

Tabelle 7.1: Mapping der Annotationen bzgl. Sichtbarkeit der Instrumente für die Datensets HeiChole und Cholec80.

Datenset	Instrumentenvektor						
	0	1	2	3	4	5	6
HeiChole	Grasper	Clipper	CoagulationInstrument	Scissors	SuctionIrrigator	SpecimenBag	Stapler
Cholec80	Grasper	Clipper	Hook	Bipolar	Scissors	Irrigator	SpecimenBag	-

Neben dem Harmonisieren von Daten und dem gemeinsamen Training, wie es für das zuvor beschriebene Experiment umgesetzt wurde, ist es insb. bei stark unterschiedlichen Daten gängig, zunächst anhand der weniger zum realen Anwendungsfall passenden Daten vorzutrainieren. Anschließend wird dann der Großteil der trainierten Gewichte eingefroren und für die letzten Schichten mit den restlichen Daten ein Finetuning durchgeführt. Dieser Ansatz wurde im Rahmen der vorliegenden Arbeit ebenfalls untersucht. Dafür wurde zunächst das komplette Transformer-Netz mit den Cholec80-Daten trainiert und anschließend das Finetuning mit den HeiChole-Daten durchgeführt. Hierfür wurden sowohl Encoder als auch Decoder eingefroren und nur noch die Fully-Connected-Layer weiter trainiert. Dies sollte die Schwächen des Cholec80-Datensatzes mit weniger Gewicht in den Trainingsprozess einfließen lassen, wobei der Vorteil der höheren Varianz und der größeren Datenquantität durch beide Datensets aber ausgenutzt werden kann.

Für alle der zuvor beschriebenen Modelle betrug der Trainings-Split der verwendeten Daten jeweils 80 Prozent für das Training und 20 Prozent für die Validierung. Weiterhin wurden, für eine bessere Vergleichbarkeit, alle mit den gleichen Hyperparametern trainiert. Die relevantesten sind nachfolgend in Tabelle 7.2 genannt.

Tabelle 7.2: Übersicht verschiedener Parameter für das Training und die Validierung der Modelle zur Phasenerkennung in Endoskopvideos.

	Parameter	Wert
Architektur	Inputlänge	10
	Dimension Embeddings	7
	# Encoder-/Decoder-Layer	1
	# Attention-Heads	1
	Dimension Feed Forward Net	128
Train	Optimizer	Adam
	Learning Rate	0,005975
	LR-Scheduler	Custom
	Weight Decay	none
	Epochen	5
	Dropout	0,1

7.3 Evaluation der trainierten Modelle zur OP-Phasenerkennung

Für die Evaluation der einzelnen in Abschnitt 7.2 beschriebenen Experimente wurde das Evaluationsscript zur Phasenerkennung genutzt, welches im Rahmen des HeiChole Benchmarks veröffentlicht wurde [14]. Dieses berechnet den F1-Score für jede Klasse sowie den Gesamt-F1-Score und die Gesamt-Genauigkeit. Die Phaseneinteilung entspricht dabei den laparoskopischen Phasen (Phasen 2 - 8) aus dem in Abschnitt 4.2.2 definierten Ablauf (vgl. Tabelle 7.3). Für die vorliegende Auswertung wurde deren Nummerierung angepasst, da hier einerseits nur diese Phasen betrachtet werden und andererseits dadurch eine Vergleichbarkeit mit den teilnehmenden Forschungsgruppen beim HeiChole Benchmark vereinfacht wird. Darüber hinaus wurde zur tiefergehenden Analyse für jedes Experiment noch eine Konfusionsmatrix und die Phasen-Zeit-Diagramme der einzelnen Evaluationsvideos erstellt (vgl. Erläuterungen in Abschnitt 2.2.4)

Tabelle 7.3: Phaseneinteilung für die Evaluation der Phasenerkennung.

Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
Vorbereitung	Dissektion Calot Dreieck	Clippen & Schneiden	Dissektion Gallenblase	Verpackung Gallenblase	Blutstillung & Spülung	Bergung Gallenblase

7.3.1 Auswertung des Modelltrainings mit Cholec80

Die Evaluationsergebnisse zur Phasenerkennung auf Basis des Trainings mit dem Cholec80-Datenset sind in Tabelle 7.4 dargestellt. Die Gesamt-Genauigkeit von 89,7 Prozent zeigt eine gute Erkennungsqualität auf den evaluierten Daten. Die Differenz von fast 15 Punkten zwischen der Genauigkeit und dem F1-Score deutet allerdings darauf hin, dass der Anteil der einzelnen Phasen nicht gleichmäßig verteilt und deren Erkennungsqualität nicht gleichermaßen hoch ist oder genauer gesagt, dass eine nicht unerhebliche Anzahl an Falscherkennungen für die einzelnen Phasen vorliegt. Dies spiegeln auch die Werte für die einzelnen Phasen wider. Hier wird deutlich, dass insb. die Dissektion des Calotschen Dreiecks mit einem F1-Score von gerundet 99 Prozent hervorragend differenziert werden kann, aber auch die Vorbereitungsphase und die Dissektion der Gallenblase zeigen mit einem F1-Score von jeweils mehr als 93 Prozent sehr hohe Werte. Dagegen fällt der F1-Score gegen Ende der Prozedur stark ab, sodass die Blutstillung & Spülung nur einen Score von 50 Prozent und die Bergung der Gallenblase sogar nur 32 Prozent aufweisen. Letztere ist im Datensatz auch am seltensten vertreten, weshalb die geringe Quantität an Trainingsdaten der Grund für die schlechte Erkennungsqualität sein könnte.

Tabelle 7.4: Evaluationsergebnis des Modelltrainings mit Cholec80.

Genauigkeit	F1-Score							
Gesamt	Gesamt	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
89,71	75,23	93,73	98,62	82,16	93,42	76,04	50,32	32,32

Die Konfusionsmatrix in Abbildung 7.2 erlaubt weitere Einblicke zur Analyse der Ergebnisse. Insb. wird deutlich, zu welchen Phasen die fehlerhaften Erkennungen zugeordnet wurden. Dabei fällt auf, dass alle falsch-positiven Vorhersagen der Vorbereitungsphase in die darauffolgende Phase zur Dissektion des Calotschen Dreiecks eingeordnet wurden. Bei genauerer Analyse des Datensets wird deutlich, dass in Phase 0 sehr häufig die Annotationen nicht dem vorgegebenen Protokoll[1] entsprechen. Dieses sieht vor, dass Phase 0 so lange stattfindet, bis das erste Instrument sichtbar ist. Danach startet Phase 1. Tatsächlich sind in Cholec80 einige Frames mit sichtbaren Instrumenten der Vorbereitungsphase zugehörig annotiert, was entsprechend zu einem fehlerhaften Modell führt. Weiterhin fällt auf, dass Frames der Phase 2 („Clippen & Schneiden“) häufig Phase 3 („Dissektion Gallenblase“) zugeordnet wurden. In den meisten Fällen betraf dies den Phasenübergang, sodass hierbei eine um einige wenige Frames zeitlich verschobene Phasenerkennung stattfand, die korrekte Reihenfolge aber bestehen blieb. Ähnliches zeigt die Analyse von Phase 4. Auch hier lagen die meisten Fehler in den Phasenübergängen. Bei der tiefergehenden Auswertung zu Phase 5 („Blutstillung & Spülung“) werden die Schwierigkeiten bei nicht sequenziellem Phasenablauf deutlich.

Wie in Abbildung 7.2 zu sehen, wird fälschlicherweise häufig Phase 3 vorhergesagt. Da diese Phase zum einen durchschnittlich zu den längsten und gleichzeitig auch zu denen mit den höchsten Blutungswahrscheinlichkeiten gehört, wird sie in der Realität verhältnismäßig häufig zur Blutstillung und Spülung unterbrochen. Die Annotationen der Ground Truth Daten bilden diesen Zusammenhang allerdings nur schlecht ab und verweilen zu häufig in der Dissektionsphase. Dadurch kommt es zu einer Vermischung der Phasen und das Modell kann rein auf Basis der Instrumentennutzung deren Unterschied nur schlecht lernen. Ein weiterer Fehlerfall ist, dass stattdessen bereits Phase 6 prädiziert wurde. Dies könnte an der Tatsache liegen, dass zwischen der Verpackung der Gallenblase und deren Bergung der Bergebeutel im Bauchraum

[1] `https://docs.google.com/document/d/1PehU09Q49fUF9HDGN0i8Kr3OOEBm17ZYDRm0bgdZIXY`, zuletzt geprüft: 25.06.2023

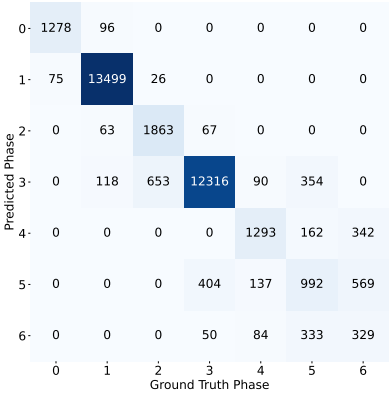


Abbildung 7.2: Konfusionsmatrix zur Evaluation der Phasenerkennung auf Basis der Cholec80-Daten.

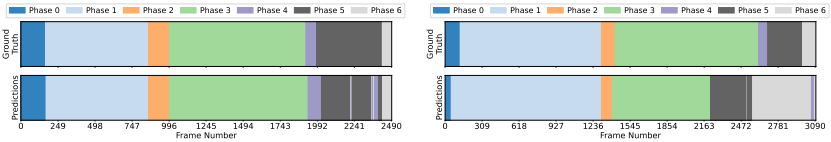
verbleibt und entsprechend durchgängig sichtbar ist. Je nach Konstellation der sichtbaren Instrumente schließt das Modell dann anstatt auf Phase 5 bereits auf Phase 6. Durch den im Konzept beschriebenen Histogramm-Mechanismus wird dieses Verhalten möglicherweise sogar verstärkt, wenn bereits mehrfach Phase 5 vorhergesagt wurde (vgl. bspw. Phasen-Zeit-Diagramm zu Video 72 in Anhang A.4.1). Für die Tatsache, dass Phase 6 nur in ca. einem Drittel der Zeit korrekt erkannt wurde, gibt es hauptsächlich zwei mögliche Erklärungen. Einerseits ist sie die durchschnittlich kürzeste Phase, weshalb entsprechend die wenigsten Trainingsdaten im Datensatz verfügbar sind. Dies erschwert ein generalisiertes Training. Die Tatsache, dass das Modell, wie in den Analysen der anderen Phasen beschrieben, oftmals Schwierigkeiten hat, die Phasenübergänge korrekt zu klassifizieren, lässt vermuten, dass die teilweise nur sehr kurze Phasendauer und das Fehlen einer nachfolgenden Phase dazu führen, dass Phase 6 teilweise gar nicht vorhergesagt wurde. Dies ist in den Phasen-Zeit-Diagrammen zu den Videos 65, 67, 69, 70, 71, 74 und 78 sichtbar (vgl. Abschnitt A.4.1). Andererseits unterscheidet sich die Phase in ihrer Instrumentennutzung nur wenig von den vorhergehenden Phasen (vgl. Tabelle 4.2). Identisch zu Phase 4 wird auch hier lediglich der Greifer und der Bergebeutel eingesetzt. Beide sind aber auch in

Phase 5 sichtbar, sodass diese Differenzierung alleine auf der Information der Instrumentennutzung nur schwer möglich ist.

Eine genauere Untersuchung der Fehlprädiktionen zeigt neben den bisherigen Erkenntnissen noch weitere häufig auftretenden Fehlerquellen, die im Folgenden näher erläutert werden. Eine der größten Herausforderungen ist, dass die Ground Truth Labels nach Auffassung des Autors dieser Arbeit nicht vollständig korrekt sind. So werden Phasenwechsel in den Ground Truth Daten häufig zu früh markiert oder bleiben, insb. bei der Phase „Blutstillung & Spülung“, ganz aus. Ersteres führt lediglich zu zeitlich um wenige Sekunden verschobene Phasenwechsel zwischen der Modellvorhersage und den Annotationen, was in einem realen Einsatz nicht problematisch wäre. Der zweite Fall tritt häufig auf, wenn während der Dissketionsphasen Blutungen einsetzen, die dann behoben werden. Tatsächlich müsste die Dissektionsphase dann unterbrochen und Phase 5 aktiv sein. Das Modell prädiziert dies auch entsprechend, die Ground Truth widerspricht hier allerdings häufig. Dies führt bei der Auswertung zu Modellfehlern, die eigentlich nicht korrekt sind. Das beschriebene Verhalten könnte auch im realen Einsatz Auswirkungen haben, bspw. wenn Statistiken über die Dauer einzelner Phasen erhoben oder die Häufigkeit oder Dauer von Blutungen erfasst werden sollen.

Eine weitere Fehlerursache, die regelmäßig auftritt, ist, dass keine Instrumente erkannt werden. Dies kann entweder auftreten, weil tatsächlich keine Instrumente im Bild sichtbar sind oder weil, bspw. aufgrund von Rauchentwicklung, die Instrumentendetektion fehlschlägt. Die Modellvorhersage ergibt in diesen Fällen meist einen Phasenwechsel. Wenn danach wieder die gleichen Instrumente im Einsatz sind, springt das Modell üblicherweise wieder zurück in die korrekte Phase, sodass hier nur kurze Abschnitte nicht korrekt erkannt werden, was im realen Einsatz wahrscheinlich wenig problematisch wäre. In den späteren Phasen des OP-Verlaufs, in denen die Phasen auch ähnlich im Instrumenteneinsatz und die Phasenübergänge ohnehin verschwommener sind, können dadurch aber ganze Phasen nicht erkannt oder vertauscht werden. Dieses Phänomen ist u. A. in den Videos 78 und 80 zu finden.

Eine weniger häufig auftretende Situation, die beim hier betrachteten Modell zu falschen Prädiktionen führt, ist ein außergewöhnlicher Einsatz von Instrumenten. Dies kann bspw. situationsbedingt oder auch aus Vorliebe des Operateurs sein. Ein Beispiel ist in Video 73 erkennbar. Hier wird in der Bergungsphase anstatt einem normalen Greifer das Bipolar zum Fassen des Bergebeutels genutzt. Da zuvor damit noch koaguliert wurde, war das in dieser Situation vermutlich weniger aufwändig, als das Instrument nochmal zu wechseln. Aufgrund der Tatsache, dass das Modell allerdings lediglich die Instrumente als Eingabe nutzt und nicht beachtet, wofür diese eingesetzt werden, hat es hier die Phase „Blutstillung & Spülung“ erkannt, in der normalerweise das Bipolar zum Einsatz kommt. Abbildung 7.3 zeigt jeweils ein Beispiel für ein gutes und ein schlechtes Ergebnis der evaluierten Phasen-Zeit-Diagramme auf dem Cholec80-Datensatz.



(a) Beispiel für gutes Evaluationsergebnis: Video 77

(b) Beispiel für schlechtes Evaluationsergebnis: Video 72

Abbildung 7.3: Phasen-Zeit-Diagramme zur Evaluation der Phasenerkennung auf Basis der Cholec80-Daten.

Generell fallen bei diesem Modell auch die Schwächen des Datensets ins Gewicht. So hat es bspw. große Schwierigkeiten Abläufe, die nicht standardgemäß und sequenziell ablaufen, zu erkennen. Dies ist u. A. in den Videos 70, 74 und 78 zu sehen, bei denen die Phase zur Blutstillung und Spülung jeweils vor dem Verpacken der Gallenblase stattfindet. Dies ist im Trainingsset nur selten vertreten und entsprechend schwer erlernbar. Weiterhin sind im Cholec80-Set kaum wiederholende Wechsel zwischen Phasen, z. B. von der Dissektion zur Blutstillung und wieder zurück, enthalten, weshalb auch diese Situationen vom damit trainierten Modell nur schwer erkannt werden. Dies tritt in den vorliegenden Auswertungen vor allem auf, wenn das Modell einen Phasenwechsel bspw. zur Blutstillung erkennt, der in den Ground Truth Daten nicht annotiert ist. Da

durch die fehlende Annotation der Wechsel zurück in die ursprüngliche Phase nicht in den Trainingsdaten enthalten ist, konnte dieses Verhalten nicht gelernt werden und das Modell erkennt die Rückkehr in die vorangegangene Phase nicht. Der gewählte Histogramm-Ansatz soll dies vermeiden, schafft es mit den vorliegenden Daten aber nicht vollständig. Schließlich fällt auf, dass insb. sehr kurze Phasen nur schwer erkannt werden. Dies kann einerseits in der geringen Menge an entsprechenden Trainingsdaten liegen, wie bereits zuvor diskutiert wurde. Andererseits ist aber auch die Modellarchitektur eine mögliche Ursache, da das sich aufbauende Histogramm die Phasenlängen berücksichtigt und eine unübliche Verteilung der Phasen hierbei Schwierigkeiten bereiten kann.

7.3.2 Auswertung des Modelltrainings mit HeiChole

Tabelle 7.5 zeigt das Evaluationsergebnis für das Training mit den Daten aus dem HeiChole-Datenset. Zunächst fällt hierbei auf, dass sowohl die Genauigkeit als auch der F1-Score deutlich um ca. 22 Punkte niedriger ausfallen als beim Modell, das mit den Cholec80-Daten trainiert wurde. Ähnlich zu den Ergebnissen des zuvor diskutierten Modells ist die Genauigkeit ebenfalls um ca. 14 Punkte höher als der F1-Score und die F1-Scores der späteren Phasen fallen im Vergleich zu den früheren Phasen deutlich ab. Davon abgesehen ist lediglich der F1-Score zu Phase 0 höher als zuvor. Bei allen anderen Phasen wurden erheblich schlechtere Ergebnisse erzielt. Die Vorbereitungsphase wurde nun nahezu fehlerfrei erkannt. Dies liegt wahrscheinlich daran, dass bei den Annotationen mehr auf die Einhaltung der Anweisungen geachtet wurde und in diesem Datenset tatsächlich keine Instrumente in dieser Phase sichtbar sind. Die Phasen 1 und 2 sind mit einem F1-Score von jeweils ca. 77 Prozent zwar prinzipiell noch gut differenzierbar, allerdings weit entfernt von den Werten des vorherigen Modells (99% & 82%). Noch deutlicher fällt der Unterschied in Phase 3 (93% vs. 54%) und Phase 4 (76% vs. 48%) aus. Besonders auffällig ist, dass Phase 5 („Blutstillung & Spülung“) nahezu gänzlich inkorrekt erkannt wurde. Die Analyse des Trainings-Datensets zeigt, dass in dieser Phase keine eindeutige

Instrumentenzuordnung möglich ist, da oftmals verschiedene Instrumente sichtbar sind. Darüber hinaus kommt das eigentlich signifikante Saug-Spühl-Rohr nicht immer zum Einsatz. Im Gegensatz zum Cholec80-Datenset ist hier die Varianz, wann die Phase stattfindet, wesentlich höher oder findet gar nicht statt. Bei der geringen Menge an Trainingsdaten konnte das Modell diese Zusammenhänge offenbar nur unzureichend lernen. Weniger deutlich ist der Absturz in Phase 6 (32% vs. 17%), was allerdings auch mit dem ohnehin schlechten Score zusammenhängen kann.

Tabelle 7.5: Evaluationsergebnis des Modelltrainings mit HeiChole.

Genauigkeit	F1-Score							
Gesamt	Gesamt	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
67,30	53,33	99,71	77,48	77,10	54,17	48,04	00,02	16,78

Die Konfusionsmatrix in Abbildung 7.4 zeigt die größte Auffälligkeit bei den Daten zur Prädiktion von Phase 1 („Dissektion Calot Dreieck“). Diese wurde signifikant häufig vom Modell vorhergesagt. Das führt dazu, dass für diese Phase zwar die Sensitivität sehr hoch ist, da allerdings sehr viele Frames fälschlicherweise dieser Phase zugeordnet wurden, ist die Präzision niedrig. Am ausgeprägtesten ist dieses Verhalten für die Phasen 5 und 6, die beide in mehr als 80 Prozent der Fälle mit Phase 1 verwechselt wurden. Die Phasen-Zeit-Diagramme in Anhang A.4.2 verdeutlichen diese Beobachtung. Auch hier ist erkennbar, dass in 3 von 5 Videos fast alle Frames, die nicht zur Vorbereitungs- oder der Clippen & Schneiden-Phase gehörten, Phase 1 zugeordnet wurden. In diesen Videos ist Phase 1 außerdem die mit Abstand längste Phase im Gesamtprozess. Ein ähnliches Bild, wenn auch mit wesentlich kleineren absoluten Zahlenwerten, zeigt sich in der Zeile zur Prädiktion von Phase 6. Diese wurde gelegentlich in allen Phasen außer Phase 0 und 1 fälschlicherweise prädiziert. Die Prädiktion der in den Trainingsdaten am häufigsten vertretenen Klasse ist ein deutliches Anzeichen von Underfitting und entsprechend zu wenigen Trainingssamples, um die komplexen Zusammenhänge adäquat erlernen zu können. In dem Fall werden nicht die relevanten Merkmale in den Trainingsdaten gelernt, sondern

hauptsächlich die Häufigkeit der auftretenden Klassen berücksichtigt. Dies ist auch daran ersichtlich, dass in allen Phasen außer 0 und 1 eine hohe Streuung in den prädizierten Phasen vorherrscht.

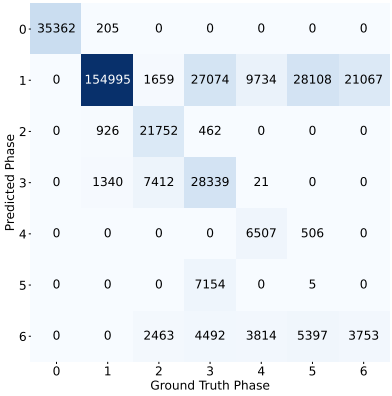
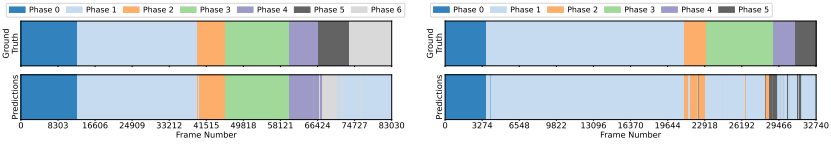


Abbildung 7.4: Konfusionsmatrix zur Evaluation der Phasenerkennung auf Basis der HeiChole-Daten.

Aufgrund der insgesamt kleinen Anzahl an Videos im Datenset, verblieben im gewählten Split nur fünf Videos als Testset für die Evaluierung der Ergebnisse. Dies ist zu wenig, um gefestigte Erkenntnisse zu wiederkehrenden Merkmalen erkennen zu können. Die Phasen-Zeit-Diagramme in Anhang A.4.2 zeigen jedoch deutlich, ähnlich wie die zugehörige Konfusionsmatrix, dass das trainierte Modell nicht in der Lage ist, alle Phasen gleichermaßen zu erkennen. Insbesondere die Phasen 3 bis 6 zeigen große Abweichungen. Auffällig ist, dass Falschprädiktionen einzelner Abschnitte offenbar großen Einfluss auf die nachfolgenden Erkennungen nehmen, sodass diese ebenfalls falsch interpretiert wurden. Abbildung 7.5 zeigt jeweils ein Beispiel für ein gutes und ein schlechtes Ergebnis der evaluierten Phasen-Zeit-Diagramme auf dem HeiChole-Datensatz.

Insgesamt scheint hier, trotz der, insb. in Bezug auf Realitätsnähe, erheblich besseren Datenqualität im Vergleich zum Cholec80-Datenset, die Menge an Trainingsdaten nicht ausreichend gewesen zu sein, um die konzeptionierte



(a) Beispiel für gutes Evaluationsergebnis: Video 20

(b) Beispiel für schlechtes Evaluationsergebnis: Video 21

Abbildung 7.5: Phasen-Zeit-Diagramme zur Evaluation der Phasenerkennung auf Basis der HeiChole-Daten.

Netzwerk-Architektur ausreichend zu trainieren. Im Gegensatz zu den Daten im Cholec80-Set, wurde das Annotationsprotokoll zum Übergang von Phase 0 zu Phase 1 sehr gut eingehalten. Folgerichtig wurde Phase 0 auch nahezu fehlerfrei erkannt. Die Anweisung gibt vor, dass Phase 1 startet, sobald ein Instrument sichtbar ist. In vielen Fällen findet allerdings unter Zuhilfenahme des Greifers zunächst eine Inspektion der Bauchhöhle statt, was eigentlich eher der Vorbereitung als der Dissektion entspricht. Das Annotationsprotokoll führt an dieser Stelle also eher zu einem verschwimmen der Phasengrenzen. Dadurch wird die Abgrenzung der Dissektion des Calotschen Dreiecks erheblich erschwert und auch in den anderen Phasen, in denen lediglich Greifer sichtbar sind, scheinen dadurch große Unsicherheiten bei der Prädiktion zu entstehen. Weiterhin fällt auf, dass Phase 5 nicht in jedem Video annotiert und damit nicht konsistent mit dem Annotationsprotokoll ist, welches besagt, dass die Blutstillung & Spülung beginnt, sobald „*der Fokus vom Bergebeutel abgewandt wird*, die Koagulation mit Strom einsetzt oder die Drainage ins Bild kommt“. Der abgewandte Fokus vom Bergebeutel ist häufig in den Daten sichtbar und zeigt sich meist in einer Inspektion des Bauchraums, ohne dass konkrete Maßnahmen durchgeführt werden. Diese Abschnitte sind in den vorhandenen Annotationen meist den Phasen 4 oder 6 zugeordnet, was ebenfalls dazu führt, dass keine klare Trennung zwischen diesen Phasen erkennbar ist und entsprechend nicht antrainiert werden kann. Dies wiederum hat zur Folge, dass in früheren Abschnitten der OP, in denen keine Instrumente sichtbar sind, vom vorliegenden Modell fälschlicherweise Phase 6 prädiziert wurde. Auch das Auftreten des Saug-Spül-Rohrs ist nicht konsequent zu Phase 5 gelabelt.

Erstaunlich ist, dass in Video 24 der Abschnitt, in dem der Stapler sichtbar ist, korrekt Phase 2 zugeordnet wurde, obwohl das Instrument in den Trainingsdaten nicht vertreten war. Es ist unklar, ob dies durch Zufall entstand oder ob das Modell andere Zusammenhänge erlernen konnte, welche die Phasenzuordnung erlauben.

Ähnlich wie beim Modell, welches mit den Cholec80-Daten trainiert wurde, waren auch hier die Phasenübergänge häufig verschoben. Wie bereits zuvor diskutiert, ist der Einfluss dieses Verhaltens allerdings marginal. Sofern keine Instrumente im Kamerabild detektiert werden konnten, wurde meist ein Phasenwechsel prädiziert. Im Gegensatz zum Cholec80-Modell wurde hier allerdings nicht immer die im sequenziellen Ablauf, wie in Tabelle 4.2 und Abbildung 4.11 dargestellt, angrenzende Phase vorhergesagt. Hier ist die größere Varianz im Datensatz bemerkbar.

7.3.3 Auswertung des Modelltrainings mit Cholec80 + HeiChole

Im folgenden Abschnitt werden die Ergebnisse zum Modell diskutiert, das mit den Daten aus Cholec80 und HeiChole zusammen trainiert wurde. Die erzielten Werte sind in Tabelle 7.6 dargestellt. Es ist direkt ersichtlich, dass das gemeinsame Training deutlich bessere Erkennungsraten ermöglicht, als das Modell, das lediglich auf den HeiChole-Daten basiert. Dabei fällt die Steigerung im F1-Score sogar größer aus als die in der Genauigkeit. Dies deutet auf eine erhöhte Präzision und somit geringere Unsicherheiten in den Vorhersagen hin. Insbesondere die zuvor besonders schlecht erkannten Phasen 4, 5 und 6 zeigen deutliche Steigerungen. Dagegen fällt der F1-Score für die Phasen 2 und 3 leicht ab. Auch in diesem Modell zeigen sich die Probleme bei der Erkennung von Phase 6 in Form des bisher mit Abstand schlechtesten F1-Scores.

Im Vergleich zum reinen Cholec80-Modell fallen die Ergebnisse sowohl in der Genauigkeit als auch im F1-Score deutlich schlechter aus (62,8% vs. 89,7%

Tabelle 7.6: Evaluationsergebnis des Modelltrainings mit Cholec80 & HeiChole, Auswertung auf gesamtem Datenset, nur Cholec80 und nur HeiChole.

Daten	Genauigkeit	F1-Score							
	Gesamt	Gesamt	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
Gesamt	72,03	62,84	99,59	82,30	69,09	48,94	70,83	46,71	22,38
Cholec80	75,30	66,20	94,78	83,64	79,26	69,37	75,19	43,52	17,64
HeiChole	71,73	62,17	99,77	82,19	68,09	45,03	70,51	46,98	22,61

bzw. 62,8% vs. 75,2%). Dies verdeutlicht die höhere Komplexität im HeiChole-Datensatz. Aber auch in diesem Vergleich ist die Differenz zwischen Genauigkeit und F1-Score im Modell mit gemeinsamem Training geringer, sodass die Unsicherheiten in den Prädiktionen insgesamt kleiner sind. Beim Analysieren der einzelnen Phasen wird deutlich, dass der F1-Score im Cholec80-Modell in allen außer der Vorbereitungsphase deutlich höher ausfällt. Besonders auffällig ist der Unterschied in Phase 3 (48,9% vs. 93,4%). Lediglich Phase 0 wurde im Cholec80-Modell schlechter erkannt. Die Problematik der Annotation dieser Phase, die wahrscheinlich für die schlechtere Performance in diesem Modell ursächlich ist, wurde bereits zuvor ausführlich diskutiert.

Tabelle 7.6 zeigt zur besseren Differenzierung neben der Gesamtauswertung auch die Erkennungsqualität des gemeinsam trainierten Modells jeweils nur auf den Cholec80- bzw. den HeiChole-Daten ausgewertet. Die Ergebnisse auf den Cholec80-Daten fallen nach wie vor schlechter aus als beim reinen Cholec80-Modell. Allerdings ist die Differenz zu diesem geringfügig kleiner als bei der gemeinsamen Auswertung. Am ausgeprägtesten ist dies in den Phasen 2 (Differenz: 2,9% vs. 13,1%) und 3 (Differenz: 24,0% vs. 44,5%) zu erkennen. Auffällig ist, dass die F1-Scores der Phasen 5 und 6 in dieser Auswerteansicht sogar weiter auseinander liegen als bei der gemeinsamen Betrachtung. Die Unterschiede zwischen den HeiChole- und den gemeinsam ausgewerteten Ergebnissen sind nur marginal, sodass hier keine relevanten Diskussionspunkte ersichtlich werden.

Die Konfusionsmatrix, basierend auf den gemeinsamen Validierungsdaten aus Cholec80 und HeiChole, in Abbildung 7.6 zeigt im Gegensatz zu derjenigen zum reinen HeiChole-Modell aus dem letzten Abschnitt (Abbildung 7.4) wieder

eine deutliche Diagonale, was darauf hindeutet, dass die einzelnen Klassen überwiegend gut vom Modell diskriminiert werden konnten. Ähnlich wie im zuvor diskutierten Modell fällt auf, dass die Sensitivität in Phase 1 nahezu optimal ist, die falsch-positiv-Rate jedoch in den eigentlichen Phasen 3 und 6 sowie in geringerem Ausmaß auch in den Phasen 2, 5 und vereinzelt in Phase 4 ebenfalls hoch ausfällt. Insgesamt sind die falsch-positiven Erkennungen dieser Klasse weniger ausgeprägt als beim reinen HeiChole-Modell. Die hierfür verantwortlichen Merkmale wurden offensichtlich aber auch in diesem Modell trainiert. Der verhältnismäßig niedrige F1-Score zu Phase 3 lässt sich anhand der Konfusionsmatrix dadurch erklären, dass einerseits während dem Verlauf dieser Phase häufig Phase 1 detektiert und zusätzlich auch Teile der Phasen 2 und 5 der Dissektion der Gallenblase zugeordnet wurden. Die Verteilung auf der horizontalen Achse ähnelt hier mehr der Verteilung beim Cholec80-Modell. Erstaunlicherweise ist die Sensitivität in Phase 4 in diesem Modell am höchsten, gleichzeitig reduziert sich jedoch die Präzision. Alle falsch-positiven Erkennungen zu Phase 5 lagen in den beiden angrenzenden Phasen. Dies deutet darauf hin, dass der zugehörige temporale Zusammenhang der Phasen korrekt trainiert wurde. Allerdings weist die Erkennung während der tatsächlich stattfindenden Phase eine breite Streuung auf. Der verbesserte F1-Score von Phase 6 ist hauptsächlich darin begründet, dass die Präzision im Vergleich zum HeiChole-Modell wesentlich höher ist. Die Sensitivität ist ähnlich ausgeprägt, wobei in diesem Modell die Streuung auf die verschiedenen Klassen sogar größer ist.

Beim Betrachten der Phasen-Zeit-Diagramme (siehe Anhang A.4.3) werden die Erkenntnisse aus der Konfusionsmatrix nochmals bestätigt. Außerdem sind im Vergleich zu den anderen Modellen nur wenige Auffälligkeiten zu diskutieren. Ein wesentlicher Unterschied zum Cholec80-Modell lässt sich darin erkennen, dass Phase 3 oftmals nahezu komplett als Phase 1 prädiziert wurde. Ein Vergleich der betroffenen Phasen-Zeit-Diagramme (Video 65, 71, 72, 76 & 79) lässt vermuten, dass dies immer dann der Fall war, wenn Phase 2 verhältnismäßig kurz ausfiel. Diese These wird bei der Untersuchung zu Video 75 noch verstärkt. Hier wurde zunächst nur ein kurzer Abschnitt als Phase 2 erkannt. Danach wurde,

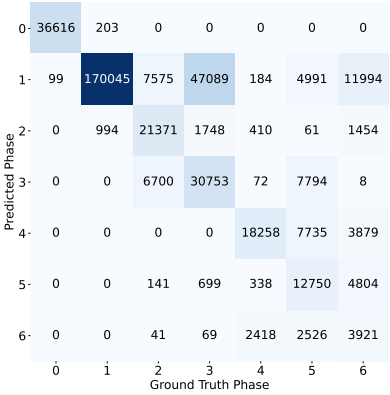
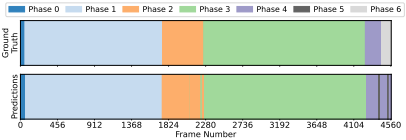


Abbildung 7.6: Konfusionsmatrix zur Evaluation der Phasenerkennung auf Basis der Cholec80- & der HeiChole-Daten.

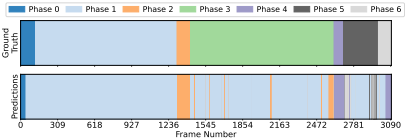
wie bei den anderen genannten Videos anstatt Phase 3 wieder Phase 1 prädiziert. Darauf folgte erneut ein (falsch-positiver) Abschnitt zu Phase 2, bevor die zeitliche Schwelle überschritten scheint und das Modell die folgenden Frames korrekt als Phase 3 erkannte. Dieses Verhalten zeigt auch, dass der konzipierte Histogramm-Ansatz das Modell wie geplant beeinflusst und die Verteilung der Phasen über die komplette Zeit der OP einfließen lässt. Jedoch funktioniert diese Lösung bisher noch nicht zufriedenstellend. Eine mögliche Erklärung dafür könnten zu wenige repräsentative Trainingsdaten sein. Eine weitere Möglichkeit könnte aber auch sein, dass der Zusammenhang, der dadurch abgebildet werden soll, zu individuell und von zu vielen Faktoren, bspw. Expertise im Team, Vorlieben des Operateurs oder bestehenden Vorerkrankungen, abhängig ist, um auf diese Weise allgemeingültig gelernt werden zu können. Letzteres könnte durch eine individuelle Anlernphase für das jeweilige Krankenhaus gelöst werden. Zur Klärung sind aber weitere Untersuchungen notwendig, die im Rahmen dieser Arbeit nicht durchgeführt werden konnten.

Interessant ist außerdem der Verlauf von Video 74, in welchem ein Großteil von Phase 3 Phase 5 („Blutstillung & Spülung“) zugeordnet wurde. Die genauere Betrachtung des Videos und der zugehörigen Annotationen zeigt, dass hier auch

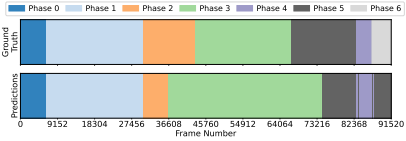
häufig sowohl das Saug-Spül-Rohr als auch das Bipolar zum Einsatz kamen. Beide repräsentieren eher Phase 5 (vgl. Abbildung 4.8b), was nach Einschätzung des Autors auch die geeignetere Annotation an diesen Stellen wäre. Die Modell-Prädiktion verweilte bei Phase 5, auch wenn längere Abschnitte ohne diese Instrumente zu sehen waren, in denen der Haken zur tatsächlichen Dissektion eingesetzt wurde. Dieses Verhalten ist an dieser Stelle nicht nachvollziehbar. Der Vergleich zwischen den HeiChole Videos zeigt zwar einige Unterschiede, allerdings können aufgrund der geringen Anzahl an Beispielvideos keine nennenswerten Trends erkannt werden. Abbildung 7.7 zeigt jeweils ein Beispiel für ein gutes und ein schlechtes Ergebnis der evaluierten Phasen-Zeit-Diagramme auf beiden Datensätzen.



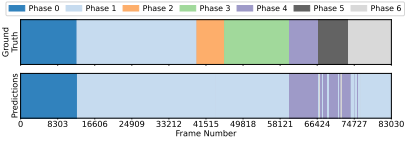
(a) Beispiel für gutes Evaluationsergebnis: Cholec80-Video 69



(b) Beispiel für schlechtes Evaluationsergebnis: Cholec80-Video 72



(c) Beispiel für gutes Evaluationsergebnis: HeiChole-Video 23



(d) Beispiel für schlechtes Evaluationsergebnis: HeiChole-Video 20

Abbildung 7.7: Phasen-Zeit-Diagramme zur Evaluation der Phasenerkennung auf Basis der Cholec80- und HeiChole-Daten.

Die Ergebnisse zeigen also gewissermaßen tatsächlich eine Mischung aus den beiden vorangehenden Modellen. Entsprechend sieht die Konfusionsmatrix an einigen Stellen wie eine Überlagerung der Konfusionsmatrizen der anderen beiden Modelle aus. Einige der Schwächen der Einzelmodelle, wie bspw. mangelnde Flexibilität bzgl. des zeitlichen Ablaufs der Phasen im Cholec80-Modell oder die schlechte Erkennbarkeit von Phase 5 im HeiChole-Modell, konnten

verbessert werden. Die Phasenübergänge und Wechsel in bereits vergangene Phasen bleiben nicht vollständig gelöste Probleme.

7.3.4 Auswertung des Modelltrainings mit Cholec80 + Finetuning mit HeiChole

Das letzte zu untersuchende Modell wurde, wie in Abschnitt 7.2 beschrieben, zunächst mit den Cholec80-Daten vortrainiert und danach ein Finetuning mit den HeiChole-Daten durchgeführt. Die Evaluation des Modells erfolgt zunächst lediglich auf Basis der HeiChole-Daten, da das Modell darauf spezialisiert sein sollte. Die zugehörigen Auswertungen bzgl. Genauigkeit und F1-Scores sind der untersten Zeile in Tabelle 7.7 zu entnehmen. Insgesamt führt dieses Vorgehen im Vergleich mit dem gemeinsam trainierten Modell wieder zu einer Verschlechterung der Erkennungsqualität, vergleichbar mit den Ergebnissen des reinen HeiChole-Modells. Besonders auffällig dabei ist einerseits, dass Phase 5 wieder einen F1-Score von 0 aufweist und dass andererseits Phase 6 in diesem Modell mit Abstand den besten F1-Score aller bisher trainierten Modelle erreicht (42,3% vs. 32,3% im Cholec80-Modell).

Tabelle 7.7: Evaluationsergebnis des Modelltrainings mit Cholec80 & Finetuning auf HeiChole, Auswertung auf gesamtem Datenset, nur Cholec80 und nur HeiChole.

Daten	Genauigkeit	F1-Score							
	Gesamt	Gesamt	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
Gesamt	71,24	56,32	98,79	82,83	57,64	64,65	44,08	04,56	41,67
Cholec80	87,04	70,28	85,17	97,46	73,94	90,76	76,58	42,77	25,30
HeiChole	69,75	54,16	99,40	81,80	56,65	58,45	40,54	00,00	42,29

Der Vergleich mit den Auswertungen lediglich auf Basis der Cholec80-Daten zeigt erneut die unterschiedliche Komplexität der Datensätze. Die Werte in der mittleren Zeile von Tabelle 7.7 bei dieser Art des Trainings liegen zwischen den Ergebnissen des reinen Cholec80-Modells und dem Modell mit gemeinsamem Training (vgl. Tabellen 7.4 & 7.6), was nicht zu erwarten war, da das Modell

sich eigentlich auf die HeiChole-Daten spezialisieren und somit weniger gut zu den Cholec80-Daten passen sollte.

Die gemeinsame Auswertung resultiert logischerweise zwischen den Werten der Einzelauswertungen (siehe oberste Zeile in Tabelle 7.7), wobei die Tendenz in Richtung der reinen HeiChole-Auswertung zeigt. Da das Finetuning auf diesen Daten stattfand, ist das Ergebnis erwartbar und auch gewünscht. Im Vergleich zum gemeinsamen Training (vgl. Tabelle 7.6) zeigt auch diese Auswertung eher eine Verschlechterung des Modells. Insb. die Phasen 2, 4 und 5 wurden schlechter erkannt (-11,5%, -26,7% und -42,1%). Die Phasen 3 und 6 konnten hingegen signifikant verbessert werden (+15,7% und +19,3%).

Zur genaueren Untersuchung der Ergebnisse bzgl. der einzelnen Phasen wird die Konfusionsmatrix für die Auswertung auf den HeiChole-Daten in Abbildung 7.8 analysiert. Auch hier zeigt sich im Vergleich zunächst eine große Ähnlichkeit zur Konfusionsmatrix zum HeiChole-Modell (vgl. Abbildung 7.4). Die größten Unterschiede liegen im starken Abfall der Präzision in den Phasen 2 und 4 sowie in der signifikanten Verbesserung sowohl der Sensitivität als auch der Präzision in Phase 6. Im Vergleich zum reinen Training mit HeiChole-Daten konnte außerdem die Präzision in Phase 1 und die Sensitivität in Phase 3 leicht gesteigert werden. In der letztgenannten Phase verschlechterte sich jedoch gleichzeitig die Präzision.

Die Phasen-Zeit-Diagramme in Anhang A.4.4 zeigen weniger Fragmentierung der Abschnitte, sodass längere zusammenhängende Phasen entstanden. Das Modell scheint also die angrenzenden Frames stärker zu gewichten, um häufige Phasenwechsel zu vermeiden. Darüber hinaus wird bei der Betrachtung der einzelnen Diagramme für die HeiChole-Videos des Testsets deutlich, dass die starke Verbesserung des F1-Scores zu Phase 6 lediglich aufgrund von Video 20 zustande kam. Hier wurde die Phase komplett erkannt, gleichzeitig ist jedoch auch ein großer Abschnitt an falsch-positiven Werten sichtbar. In den restlichen vier Videos wurde Phase 6 nie erkannt. Da die Abschnitte alle verhältnismäßig kurz sind, fallen sie aber weniger stark ins Gewicht, weshalb dennoch der recht hohe Score erzielt werden konnte. Zusätzlich fällt auf, dass es, insb. auch bei den

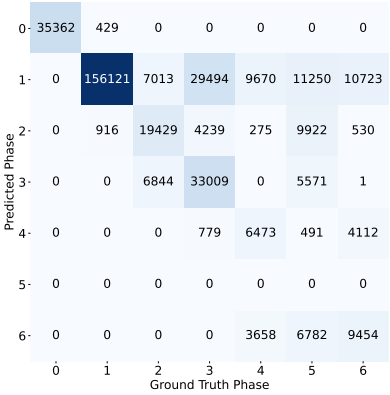
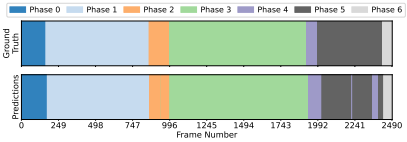


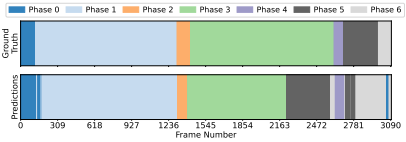
Abbildung 7.8: Konfusionsmatrix zur Evaluation der Phasenerkennung auf Basis der Cholec80-Daten mit Finetuning auf den HeiChole-Daten.

Fehldetektionen, Ähnlichkeiten in den Diagrammen der drei Modelle gibt. So wurde jeweils in den Videos 21, 22 und 24 ein Großteil der zweiten OP-Hälfte fälschlicherweise Phase 1 zugeordnet. Des Weiteren ist das Muster zu Phase 3 in Video 23 zu erwähnen. Hier fand bei den beiden Modellen, bei denen Cholec80-Daten im Training verwendet wurden, ein deutlich früherer Wechsel von Phase 2 zu Phase 3 statt, als es die Ground Truth vorgibt. Abbildung 7.9 zeigt jeweils ein Beispiel für ein gutes und ein schlechtes Ergebnis der evaluierten Phasen-Zeit-Diagramme auf beiden Datensätzen.

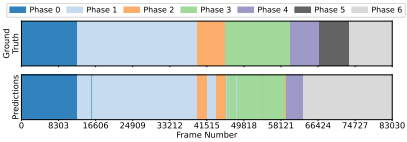
Insgesamt hat das Vortraining mit den Cholec80-Daten und anschließendem Finetuning mit den HeiChole-Daten nur für einzelne Phasen zu Verbesserungen geführt. Das Gesamtergebnis über alle Phasen hinweg hat sich allerdings leicht verschlechtert. Unklar ist, ob sich dies mit mehr Trainingsdaten entweder im Vortraining oder in der Finetuning-Phase verbessern würde. Eine Beobachtung, die für eine potenzielle Verbesserung spricht, ist die Tatsache, dass das gewünschte Verhalten durch das gewählte Vorgehen ansatzweise durchaus eintritt. So konnten Stärken der einzelnen Datensets scheinbar verstärkt in das Modell einfließen. Bspw. tritt das Phänomen der häufigen falsch-positiven Erkennung in Phase 1 aus dem gemeinsam trainierten Modell zumindest in den Cholec80-Daten nicht auf



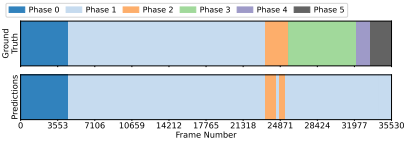
(a) Beispiel für gutes Evaluationsergebnis: Cholec80-Video 77



(b) Beispiel für schlechtes Evaluationsergebnis: Cholec80-Video 72



(c) Beispiel für gutes Evaluationsergebnis: HeiChole-Video 20



(d) Beispiel für schlechtes Evaluationsergebnis: HeiChole-Video 22

Abbildung 7.9: Phasen-Zeit-Diagramme zur Evaluation der Phasenerkennung auf Basis der Cholec80-Daten und Finetuning auf den HeiChole-Daten.

(vgl. Phasen-Zeit-Diagramme in Anhang A.4.4). Trotzdem scheint das Modell eine erhöhte Flexibilität in Bezug auf den Ablauf der Phasen beizubehalten.

7.4 Diskussion & Fazit zum Teilsystem zur OP-Phasenerkennung in endoskopischen Videos

Zusammenfassend konnte gezeigt werden, dass der konzipierte Ansatz eines transformerbasierten Modells mit Phasenhistogramm zur Analyse der Instrumentennutzung geeignet ist, um unterschiedliche Phasen innerhalb der Cholezystektomie zu diskriminieren (Anforderung [*AF-01*]).

Die Herausforderung dabei ist, vorhandene Beispieldaten so zu konsolidieren, dass daraus ein einheitliches, nutzbares Trainingsdatenset entsteht. Bestehende Datensets unterscheiden sich teilweise stark, insb. in der Qualität der Daten oder deren Annotationen. Dies hat zur Folge, dass zum aktuellen Stand der Technik, stets Kompromisse akzeptiert werden müssen, wenn verschiedene

Datensets kombiniert werden sollen, um eine ausreichende Quantität und Diversität innerhalb der Trainingsdaten zu erhalten. Das Erzeugen eigener Daten, die alle benötigten Anforderungen erfüllen, ist einerseits aus ressourcentechnischer Sicht, aber auch aufgrund mangelnder technischer Voraussetzungen in den Krankenhäusern kaum möglich. Letzteres wird u. A. auch in [111] diskutiert.

Durch das Training des konzipierten Modells auf unterschiedlichen Datengrundlagen konnten verschiedene Einflussquellen auf die Erkennungsqualität untersucht werden. Dabei wurden diverse, unabhängig von der Datenqualität bestehen Herausforderungen, die typisch für den realen Ablauf von OPs sind und auch bei qualitativ hochwertigen Daten auftreten können identifiziert. Beispiele hierfür sind:

1. Instrumente nicht sichtbar oder nicht erkennbar, bspw. aufgrund von Verdeckung, Verschmutzung, Bewegungsunschärfe oder Kameraschwenk zur Umfeldinspektion
2. Endoskopkamera außerhalb des Körpers, bspw. zur Reinigung
3. Untypischer Instrumenteneinsatz
4. Nicht eindeutige Phasenwechsel bzw. Phasenzugehörigkeit, bspw. durch Blutstillung während Dissektion

Bei den Trainings, die jeweils nur auf einem der beiden genutzten Datensets durchgeführt wurden, stellte sich die geringere Komplexität innerhalb der Cholec80-Daten als positiv heraus. Hierbei konnten mit einer Genauigkeit von nahezu 90 Prozent und einem F1-Score von ca. 75 Prozent hohe Erkennungsraten erzielt werden. Insb. die Genauigkeit ist vergleichbar mit dem Stand der Technik (vgl. Abschnitt 3.3). Der F1-Score fällt etwas schlechter aus, dafür ist die hier vorgestellte Modellarchitektur weniger komplex, da lediglich die erkannten Instrumente als Eingabedaten verwendet werden. Der starke Abfall der Ergebnisse beim Training mit den HeiChole-Daten spiegelt auch die Erfahrungen aus dem Stand der Technik wider. Die erzielten Ergebnisse sind dabei vergleichbar mit den in [159] präsentierten. Im direkten Vergleich mit den

veröffentlichten Resultaten in [159] befindet sich das hier vorgestellte Modell mit einem durchschnittlichen F1-Score von 53,3 Prozent im Mittelfeld (Platz 6 von 11). Beim Betrachten der einzelnen Phasen wird deutlich, dass vor allem die Phasen 5 und 6 für die Differenz verantwortlich sind. Bei den restlichen Phasen konnten vergleichbare Scores erreicht oder die bisher veröffentlichten sogar übertroffen werden.

Beim gemeinsamen Training fällt das Resultat für die Auswertung auf den Cholec80-Daten etwas ab. Für die HeiChole-Daten konnte aber eine deutliche Steigerung erreicht werden, sodass hier im Vergleich zu [159] noch eines der dort genannten Teams überholt werden konnte. Die Differenz im durchschnittlichen F1-Score betrug hierbei im Vergleich zum besten Modell lediglich 6,61 Prozentpunkte, zum zweiten Platz sogar nur 3,21 Prozentpunkte. Dies verdeutlicht auch das enge Feld unter den vorderen Teams. Der Abstand zwischen dem besten im Paper genannten Modell und Platz 4 beträgt lediglich 5,18 Prozentpunkte. Bei den eigenen Resultaten spielen die letzten Phasen des Gesamtprozesses eine entscheidende Rolle. Phase 6 wird zwar nach wie vor erheblich schlechter erkannt als in den genannten Arbeiten, allerdings mit wesentlich geringerem der Abstand. Darüber hinaus ist Phase 5 auf einem ähnlichen Niveau und Phase 4 kann sogar deutlich besser prädiziert werden als in veröffentlichten Daten der Vergleichsarbeiten.

Der gewünschte Effekt beim Finetuning, das Modell auf die spezifischen Eigenschaften der HeiChole-Daten zu optimieren und dennoch die erhöhte Datenquantität durch das Vortrainieren mit Cholec80 zu bewahren, ist ausgeblieben. Eine mögliche Erklärung dafür ist die Art und Weise des gewählten Finetunings. Es wurden lediglich die Fully-Connected-Layer zur Bestimmung der Phase weiter trainiert. Die zugrundeliegenden Transformerschichten wurden eingefroren. Möglicherweise muss hier beim Finetuning in den tieferen Schichten angesetzt werden.

Der Vergleich mit den Ergebnissen aus [159] ist dabei nicht komplett konsistent, da in der offiziellen Challenge das komplette veröffentlichte Datenset zum Training genutzt wurde und neun weitere, nicht öffentliche Videos, zur Evaluation

verfügbar waren. Insofern werden bei identischen Trainings- und Evaluationsdaten leicht abweichende Ergebnisse erwartet. Das Plus an Daten könnte dabei tendenziell zu Verbesserungen führen. Andererseits könnten die unbekannten Evaluationsdaten auch eine Verschlechterung hervorrufen.

Insgesamt konnten mit den dargelegten Untersuchungen sowohl auf dem Cholec80- als auch auf dem HeiCHole-Datensatz Resultate nahe dem Stand der Technik bei geringerer Modellkomplexität erreicht werden (Anforderung [*S-AF-02*]). Um diese Ergebnisse noch zu übertreffen, sind verschiedene Optimierungsansätze denkbar, die teilweise auch bereits bei der Ergebnisdiskussion erörtert wurden. Ein zusätzlicher Punkt, der bei der Betrachtung der Phasen-Zeit-Diagramme auffällt, ist, dass die Phasenübergänge häufig mit Unsicherheit behaftet sind und dadurch eine Art Jitter, also häufiger Wechsel der Phasenprädiktionen, entsteht. Dies könnte durch eine Glättung der Vorhersagen, bspw. in Form eines *gleitenden Mittelwerts*, gelöst werden.

Andere Möglichkeiten zur Optimierung benötigen grundsätzlichere Änderungen. So könnte eine feinere Phaseneinteilung oder weitere Unterteilung der Instrumentenklassen eine Erkennung auf Basis der Instrumentennutzung vereinfachen.

Außerdem wurde in der vorliegenden Arbeit lediglich ein transformerbasierter Ansatz evaluiert. Die zugrundeliegenden Überlegungen, dass lediglich die Instrumentennutzung in Kombination mit dem Histogramm der zuvor erkannten Phasen zur Prädiktion der nachfolgenden Phasen genutzt wird, können allerdings auch mit anderen Architekturen, wie bspw. TCNs umgesetzt werden.

Eine Einschränkung der hier dargelegten Ergebnisse ergibt sich daraus, dass bisher ausschließlich mit Ground Truth Daten der sichtbaren Instrumente gearbeitet wurde. Es ist zu erwarten, dass ein realer Objekterkenner eine signifikante Unsicherheit einbringt und dadurch die Phasenerkennung mit den aktuell trainierten Modellen negativ beeinflusst. Gleichzeitig ist zu erwarten, dass ein Einbezug der Ungenauigkeiten der Instrumentenerkennung in das Modelltraining zur Phasenerkennung ein solches Verhalten lernen und damit den Einfluss minimieren sollte.

8 Zusammenfassung und Ausblick

In diesem Kapitel werden die wesentlichen Arbeitsschritte und die daraus folgenden Resultate und Erkenntnisse dieser Arbeit rekapituliert. Im Anschluss wird ein Ausblick zu konsequenten Folgearbeiten gegeben und wie die dargestellten Entwicklungen im praktischen Arbeitsalltag innerhalb des OP-Umfelds Einzug finden können.

8.1 Zusammenfassung

Motiviert wurde die vorliegende Arbeit zunächst durch Herausforderungen der OP-Logistik und speziell des OP-Managements zur Optimierung der Planung von Operationen, was gleichermaßen zu höherem wirtschaftlichem Erfolg, verbesserter Mitarbeitendenzufriedenheit und letztendlich zu gesteigertem Patientenwohl führt. Daraus folgte als Ziel dieser Arbeit die Entwicklung einer akkuraten und robusten Erfassung komplexer Aktivitäten während laufender Operationen als wesentlicher Bestandteil der zuvor genannten Optimierungen. Die Konzeption eines solchen Systems erfolgte beispielhaft anhand der laparoskopischen Cholezystektomie als weit verbreitetes und in der Wissenschaft gut untersuchtes Operationsverfahren. Hierfür wurden zunächst die notwendigen Grundlagen zu Abläufen und Prozessen im OP-Umfeld und spezifische medizinische Grundlagen zur laparoskopischen Cholezystektomie skizziert. Auf Basis dieser Erkenntnisse, der technischen Grundlagen zur Kontexterfassung

und dem aktuellen Stand der Technik und Wissenschaft auf den angrenzenden Themenfeldern wurde eine technische Analyse des Prozesses durchgeführt und in konkrete Systemanforderungen überführt. Darauf aufbauend wurde schließlich ein Systemkonzept entwickelt, das einen modularen, multimodalen Multitasking-Ansatz verfolgt, indem verschiedene Teilsysteme eingesetzt werden, um spezifische Aspekte des OP-Ablaufs zu analysieren. Eine Risikoanalyse beleuchtete schließlich kritisch das Gesamtkonzept dieser Arbeit bzgl. verschiedener Faktoren.

Die grundlegende Architektur des Gesamtkonzeptes (vgl. Abschnitt 4.3) adressiert dabei die Anforderungen [*AF-02*], [*AF-08*], [*AF-13*], [*S-AF-01*] sowie [*S-AF-05*].

Für die Umsetzung des erarbeiteten Konzepts wurden drei eigenständige Teilsysteme entwickelt:

1. System zur Beobachtung des OP-Saals mittels Deckenkamera für die Erfassung von Personen und individueller sowie kollektiver Bewegungsabläufe
2. System zur Beobachtung des Instrumententisches mittels Deckenkamera für die Erfassung der Instrumentennutzung
3. System zur Beobachtung des Bauchinnenraumes mittels Endoskopkamera für die Erfassung von OP-Phasen

Für das Teilsystem zur Aktivitätserkennung im OP-Saal konnte aufgezeigt werden, dass mithilfe von Methoden der Posenerkennung verschiedene Bewegungsarten analysiert werden können, welche Rückschlüsse auf unterschiedliche Aktivitätslevel innerhalb des OP-Saals zulassen. Im Rahmen dieser Implementierungen wurde außerdem der Einfluss medizinischer Kleidung aufgrund von Verdeckung maßgeblicher Diskriminierungsmerkmale auf bestehende Methoden zur Re-Identifikation von Personen untersucht. Dabei wurden erwartungsgemäß starke Einbußen bei der Genauigkeit festgestellt und mögliche Mitigationsstrategien diskutiert. Das Teilsystem trägt zur Erfüllung der Anforderungen [*AF-03*],

[*AF-04*], [*AF-05*], [*AF-09*], [*AF-10*], [*AF-11*], [*AF-12*], [*AF-13*] und [*S-AF-03*] bei.

Im Rahmen der Umsetzung des Teilsystems zur Beobachtung des Instrumententisches wurde die Tauglichkeit verschiedener moderner Objekterkennungsverfahren für die Detektion der zuvor definierten Instrumentenklassen nachgewiesen. Insgesamt konnten hierbei nur marginale Qualitätsunterschiede zwischen den getesteten Modellen festgestellt werden. Eine Evaluation im realen Umfeld steht allerdings aufgrund u. A. pandemiebedingter Einschränkungen noch aus. Das Teilsystem erfüllt dabei die Anforderungen [*AF-02*], [*AF-06*], [*AF-07*], [*AF-09*], [*AF-10*], [*AF-11*], [*AF-12*] und [*S-AF-04*].

Die Untersuchungen zum Teilsystem zur Analyse des Endoskopbildes ergaben, dass eine moderne, transformerbasierte Modellarchitektur in der Lage ist, lediglich mittels der Informationen zur Instrumentennutzung OP-Phasen zu erkennen. Die Ergebnisse lagen dabei in einem mit dem aktuellen Stand der Technik vergleichbaren Bereich, wobei die Modellkomplexität und der damit verbundene Ressourcenbedarf deutlich reduziert wurde. Mit diesem Teilsystem wurden die Anforderungen [*AF-01*], [*AF-03*], [*AF-09*], [*AF-10*], [*AF-11*], [*AF-13*] und [*S-AF-02*] erfüllt.

Die größte Herausforderung bei der Umsetzung der konzipierten Ideen stellte der Mangel an adäquaten Trainingsdaten dar. Dies hatte zur Folge, dass viel Aufwand in deren Beschaffung und Erstellung fließen musste. Dennoch konnte während des Bearbeitungszeitraums der vorliegenden Arbeit kein einzelner konsistenter Datensatz erstellt werden, der den Anforderungen des Gesamtkonzeptes mit allen geplanten Teilsystemen genügt. Aus diesem Grund wurden die drei Teilsysteme separat betrachtet und evaluiert. Eine Auswertung hinsichtlich der realen Aussagekraft des Gesamtsystems für die Workflowanalyse während der laparoskopischen Cholezystektomie konnte somit nicht erfolgen. Die Annahmen aus der Konzeptphase und die erzielten Resultate der Teilsysteme deuten aber auf einen signifikanten Mehrwert durch die finale Umsetzung des erarbeiteten Konzeptes hin.

Insgesamt konnten durch die vorliegende Arbeit folgende Erfolge zu verschiedenen wissenschaftlichen Fragestellungen beigetragen werden:

1. Entwicklung einer Methode zur Aktivitätsanalyse einzelner Individuen und Personengruppen auf Basis von Videoaufnahmen aus der Vogelperspektive.
2. Untersuchung von Einflussfaktoren medizinischer Kleidung auf die Re-Identifizierbarkeit von Personen in Videobildern und entsprechender Mitigationsstrategien.
3. Adaption moderner Objekterkennungsverfahren an ein medizinisches Anwendungsfeld zur Detektion von OP-Instrumenten in Kamerabildern.
4. Entwicklung einer OP-Phasenerkennung auf Basis der Instrumentennutzung und des bisherigen OP-Verlaufs mittels einer Transformer-Architektur.

Konkret konnte die vorliegende Arbeit auch einen Beitrag zu einigen der in [111] genannten Ziele für zukünftige Arbeiten im Bereich der SDS leisten. So trägt die Homogenisierung der Cholec80- und der HeiChole-Daten in Kapitel 7 zu Ziel 2.1 bei. Das Training und die Evaluation auf verschiedenen Datensets aus unterschiedlichen Quellen und Domänen führt üblicherweise zu erhöhter Generalisierbarkeit und unterstützt damit Ziel 3.1. Der Fokus auf abgesteckte Teilprobleme durch das separate Training der einzelnen Teilsysteme (vgl. Kapitel 4) sowie die Verringerung der Modellkomplexität wirken in Ziel 3.3 ein. Ziel 3.4 wird durch den Histogrammansatz der Phasenerkennung zum Umgang mit der ungleichen Verteilung der Frames, die konkreten Phasen zugehören und jenen, die Phasenübergänge repräsentieren, adressiert. Die Tatsache, dass das Gesamtkonzept auf einer ausgiebigen Prozessanalyse basiert (vgl. Kapitel 4), wirkt mit in Ziel 3.5 ein und zu Ziel 3.6 wird beigetragen, indem echtzeitfähige Objekterkenner eingesetzt werden (vgl. Kapitel 6) und die Modellkomplexität für die Phasenerkennung gering gehalten wird.

Insgesamt konnte ein wertvoller Beitrag zur Beantwortung der anfangs gestellten Forschungsfrage bzgl. der Erfassung und Analyse komplexer Handlungsabläufe geleistet werden.

8.2 Ausblick

Ziel zukünftiger Arbeiten sollte es sein, die noch offene Gesamtsystemintegration zu verwirklichen. Um die dafür notwendigen Daten zu gewinnen, wurde bereits ein Aufnahmesystem zur systematischen Datenakquise in einem Partnerkrankenhaus installiert, wovon zeitnah nutzbare Datensätze erwartet werden. Da es sich dabei um eine urologische Klinik handelt, müssen die auf die laparoskopische Cholezystektomie spezifizierten Konzepte jedoch an die dann verfügbaren Operationstypen angepasst werden. Dies erhöht auch gleichzeitig die Generalisierbarkeit des Systems.

Im Zuge dessen muss das Konzept zur Umsetzung als multimodales Multitask-Netzwerk für die praktische Umsetzung weiter ausgearbeitet werden. Hu and Singh adressieren in [72] ähnliche Herausforderungen, die zur Problemlösung herangezogen werden können.

Diverse Optimierungsmöglichkeiten der umgesetzten Teilsysteme wurden in den entsprechenden Kapiteln bereits ausführlich diskutiert. Einen zentralen Faktor bildet dabei die Erprobung mittels realer Anwendungsdaten und der daraus folgenden Ableitung von Maßnahmen zur Steigerung der Erkennungsqualität und Robustheit. Weiterhin sind der Einsatz anderer Modellarchitekturen, wie bspw. TCNs zur Phasenerkennung, die Nutzung von Pseudolabels zur Verminderung des Problems der Annotation von Trainingsdaten oder die Anwendung weiterer Auswertemethoden und Filter zur Bewegungsanalyse mögliche Ansatzpunkte.

Ein verhältnismäßig direkter Einsatzzweck des Teilsystems zur Instrumentendetektion ist die Kontrollmöglichkeit, ob alle in den Körper eingeführten Materialien vor dem OP-Ende auch wieder entfernt wurden. Im Körper vergessenes

OP-Material ist nach wie vor Ursache vieler Behandlungsfehler und daraus folgender Gesundheitsschäden [125].

Der Blick in die weitere Zukunft verspricht diverse Einsatzmöglichkeiten für das hier vorgestellte Gesamtsystem und auch dessen Teilkomponenten. Neben der vielfach diskutierten dynamischen Planungsoptimierung können die Informationen aus einem solchen Erkennungssystem auch zur generellen Prozessanalyse und besserem Verständnis typischer Abläufe genutzt werden. Damit kann eine weitere Standardisierung von OP-Abläufen und demzufolge perspektivisch bessere OP-Resultate für Patienten durch höhere Qualitätsstandards erreicht werden. Gleichzeitig kann durch eine solche erweiterte Prozesskenntnis auch Vergleichbarkeit und demnach auch ein Benchmarking bspw. von OP-Teams zum Aufdecken von Optimierungsmöglichkeiten und Schulungsbedarfen umgesetzt werden. Dadurch kann sowohl das Patientenwohl verbessert, als auch die Effizienz im Krankenhaus gesteigert werden. Weiterhin birgt ein solches System das Potenzial, bisher unbekannte oder wenig bedachte Zusammenhänge aufzudecken und auszunutzen. So können z. B. durch Einbezug zusätzlicher Informationen genauere Abhängigkeiten der Aufwachzeit von der Narkoselänge oder vom Gewicht oder Alter der Patienten abgeleitet und somit die Narkose individueller auf den Patienten abgestimmt und der präoperative Betreuungsbedarf besser geplant werden. Außerdem können gesammelte Informationen eines integrierten OP-Saals helfen, den Narkoseprozess detaillierter zu überwachen und dadurch Spontanatmung oder ähnliche für den Anästhesisten relevante Phänomene automatisiert zu erkennen.

Neben den Aspekten, die den konkreten OP-Ablauf betreffen, kann mit den vorgestellten Methoden auch ein Beitrag zur Sicherheit und Gesundheitsförderung der Mitarbeitenden geleistet werden. So können durch die Analyse der Skelettstrukturen aus dem Teilsystem zur Raumbeobachtung auch Rückschlüsse auf Ergonomieaspekte getroffen werden. Achsveränderungen im Körperskelett können bspw. auf schlechte Körperhaltung und dementsprechend auf ein nicht optimales Raumdesign, bspw. bzgl. Tischeinstellung oder Monitorplatzierung,

hinweisen. Weitere Beispiele zur Steigerung des Personalschutzes sind die Detektion des Abstands von Personen zu Röntgengeräten oder die Erkennung nicht getragener Schutzkleidung.

Als letzten anzuführenden Punkt kann ein solches Erkennungssystem als Grundlage für weitere Assistenzsysteme, bspw. aus der Robotik dienen, da für diese üblicherweise bestimmte Informationen wie die Lokalisierung von Instrumenten sichergestellt sein müssen, um die eigentlichen Assistenzfunktionen erfüllen zu können.

Dies ist nur ein Auszug möglicher Erweiterungen und Einsatzzwecke des in dieser Arbeit entwickelten Systems und zeigt dessen enormes Potenzial. Der Einsatz von KI wird zukünftig auch in der Chirurgie immer größere Bedeutung gewinnen und fester Bestandteil des Arbeitsalltags sein. Die hier vorgestellte Arbeit leistet einen Beitrag zur Bewältigung dieser Aufgabe.

A Anhang

A.1 Gesamtsystem & Hardwaredetails

A.1.1 Details zu eingesetzter Hardware

Tabelle A.1: Hardwaredetails der eingesetzten Trainings- und Evaluationssysteme

Parameter	System 1	System 2
CPU	Intel Core i5-8600K	Intel Core i7-12700K
Kerne	6	12
Taktfrequenz	3,6 GHz	3,6 GHz
RAM	32 GB	128 GB
GPU	NVIDIA GTX 1080Ti	NVIDIA RTX A6000
GPU RAM	12 GB	48 GB

Tabelle A.2: Technische Daten der Kamera [22]

Parameter	Wert
Hersteller	Canon
Modell	VB-H45
Bildsensor	1/3-Zoll-CMOS-Sensor (mit Primärfarbenfilter)
Objektiv	20-fach optisches Zoomobjektiv mit Autofokus
Max. Auflösung	1920×1080 Pixel
Max. Bildfrequenz	30 FPS
Lichtstärke	F1.6 (Weitwinkel) – F3.5 (Teleobjektivweite)
Min. Beleuchtungsstärke	Tagmodus (Farbe): 0,05 Lux
	Nachtmodus (SW): 0,003 Lux
Schwenkwinkelbereich	340°
Neigungswinkelbereich	100°

A.1.2 Konzept Gesamtsystem

Abbildung A.1: Legende der Ablaufdiagramme des Gesamtsystems

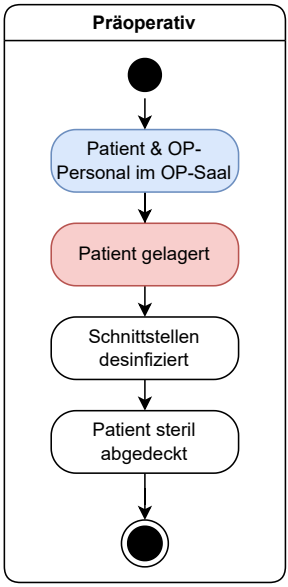


Abbildung A.2: Ablaufdiagramm der Phase „Präoperativ“

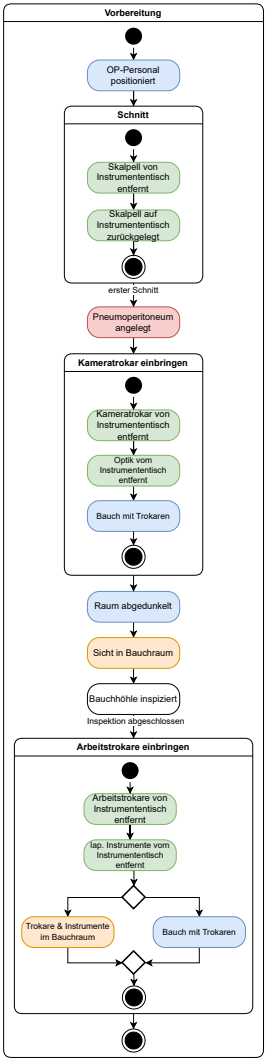


Abbildung A.3: Ablaufdiagramm der Phase „Vorbereitung“

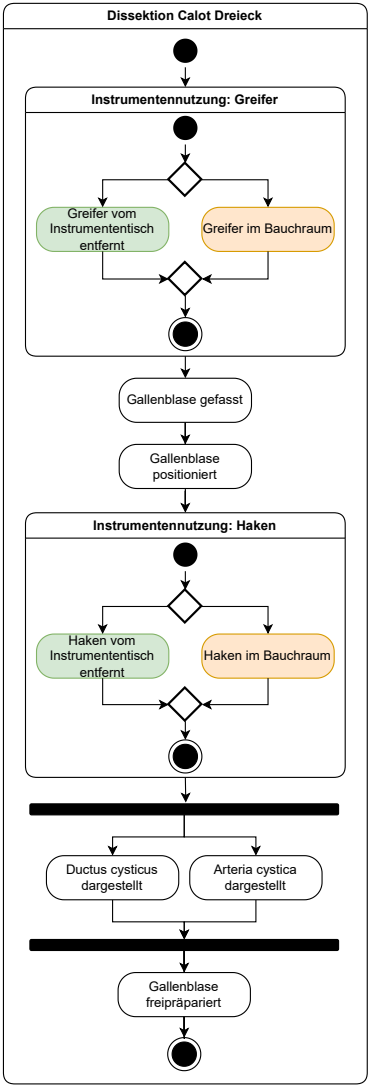


Abbildung A.4: Ablaufdiagramm der Phase „Dissektion Calot Dreieck“

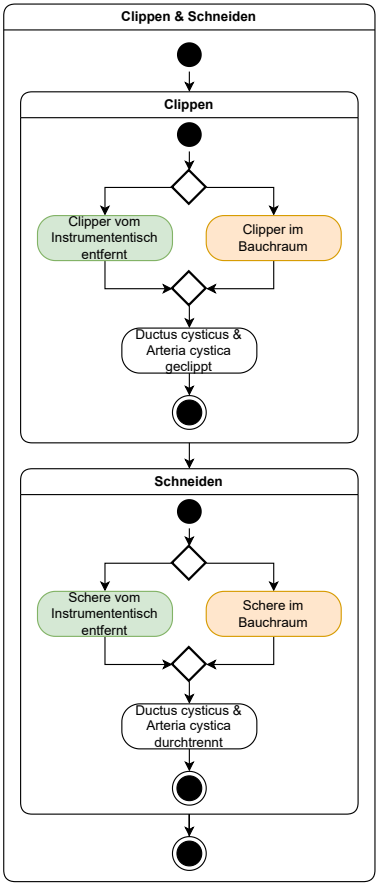


Abbildung A.5: Ablaufdiagramm der Phase „Clippen & Schneiden“

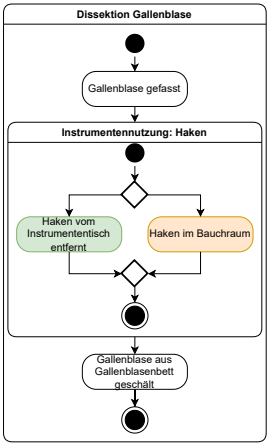


Abbildung A.6: Ablaufdiagramm der Phase „Dissektion Gallenblase“

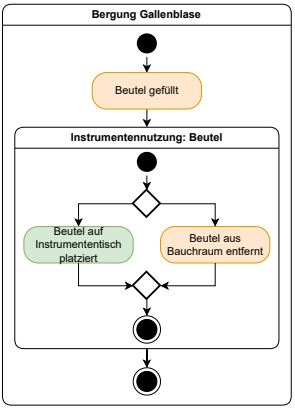


Abbildung A.7: Ablaufdiagramm der Phase „Bergung Gallenblase“

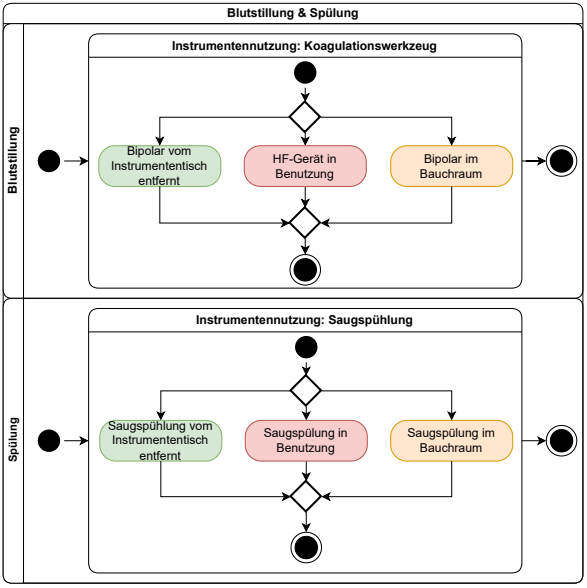


Abbildung A.8: Ablaufdiagramm der Phase „Blutstillung & Spülung“

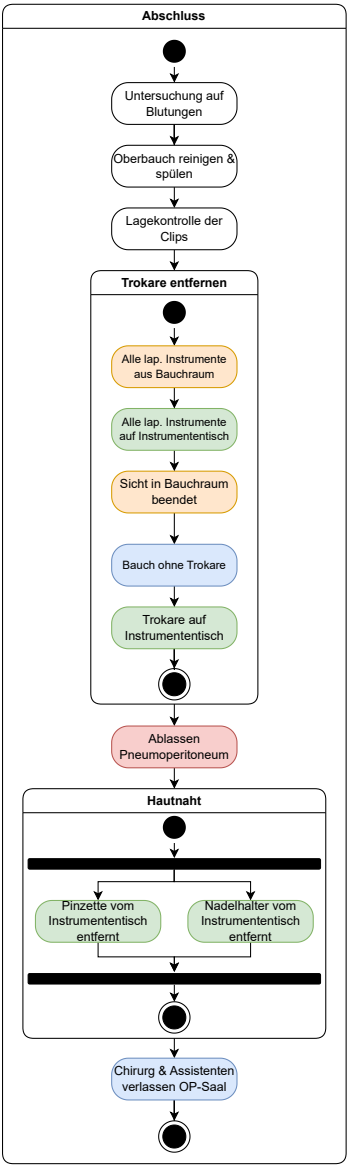


Abbildung A.9: Ablaufdiagramm der Phase „Abschluss“

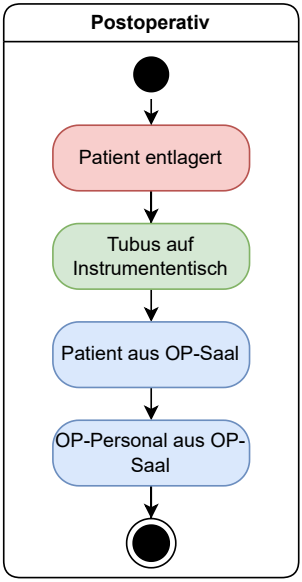


Abbildung A.10: Ablaufdiagramm der Phase „Postoperativ“

A.2 Details zu Re-ID-Untersuchungen

A.2.1 Probanden & Stationen zur Datenaufzeichnung

Tabelle A.3: Beschreibung der Probanden im Re-ID-Datensatz

Person (ID)	Beschreibung Kleidung (OP0)	Haare	Haarnetzfarbe (OP2)
0	roter Pullover, schwarze Hose, dunkelgraue Sportschuhe	braun, lang, offen	pink
1	dunkelblaues T-Shirt, schwarze Jeans, schwarze Nike-Schuhe	hellbraun, kurz	schwarz
2	blaues T-Shirt mit weißen Streifen, schwarze Hose, hellgraue Nike-Schuhe, Brille	braun, lang, Dutt	gelb
3	schwarz/weiß gestreiftes T-Shirt, schwarze Hose, schwarze Sportschuhe	dunkelblond, lang, Zopf	orange/rot
4	dunkelblaues T-Shirt, dunkelblaue Hose, braune Schuhe, Brille	braun, kurz, Locken	hellblau/violett
5	schwarzer Pullover mit Aufdruck, blaue Jeans, schwarze Schuhe	hellbraun, kurz	grün

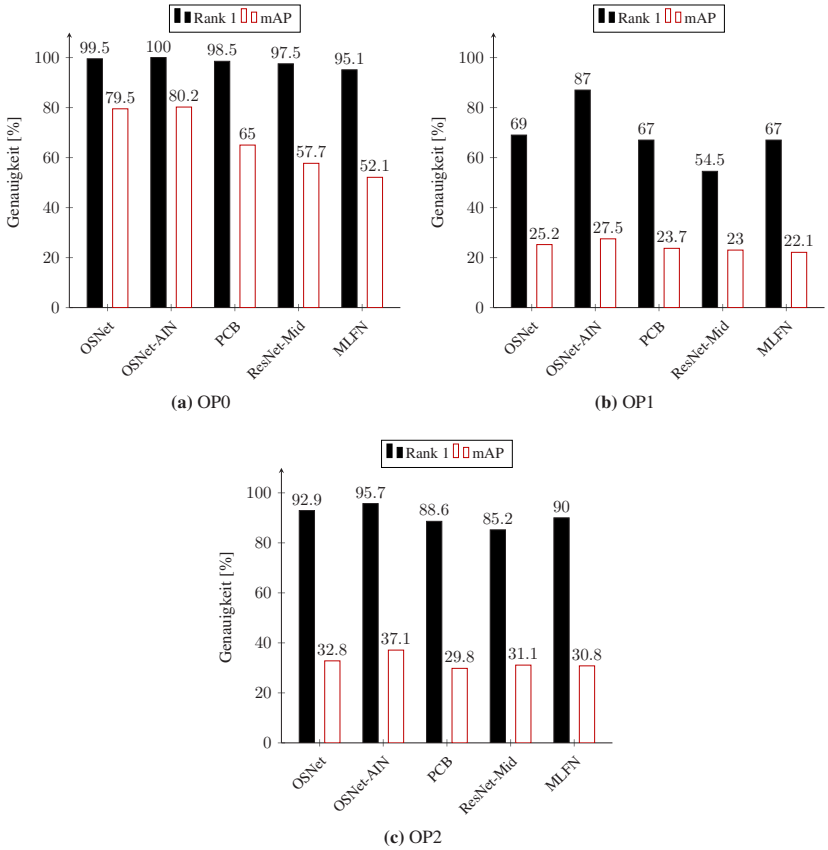
Tabelle A.4: Beschreibung der Stationen im Re-ID-Datensatz

Station	Stationsname	Kurzbeschreibung
1	OP-Tisch	Mit medizinischen Greifzangen werden Papierschnipsel in verschiedene Boxen sortiert.
2	Medikamente sortieren	Aus einem Behälter müssen verschiedenfarbige Murmeln der Farbe nach in Becher sortiert werden.
3	OP-Tisch + Transport	Mit medizinischen Greifzangen werden Papierschnipsel in verschiedene Boxen sortiert. Ein Tablett mit medizinischen Instrumenten muss an eine andere Station gebracht werden.
4	Warten	Die Station befindet sich außerhalb des Sichtbereichs der Kamera. Es muss 1,5 Minuten gewartet werden. Anschließend geht die Person weiter zur nächsten Station.
5	Vorbereitung	Ein Becheranordnung aus 5×5 Bechern muss entsprechend einer Mustervorlage mit farbigen Steinen befüllt werden.
6	Dokumentation	Auf einem Tablett wird ein Minispiel gelöst, um einen Dokumentationsvorgang zu simulieren.
7	Transport	Die Station befindet sich außerhalb des Sichtbereichs der Kamera. Ein Tablett mit medizinischen Instrumenten muss an eine andere Station gebracht werden.
8	Instrumente vorbereiten	Medizinische Instrumente müssen von einem Tablett auf ein anderes umgelegt werden.

A.2.2 Re-ID-Auswertung

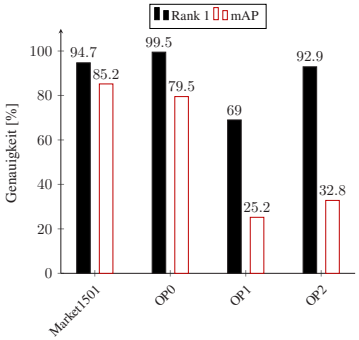
Tabelle A.5: Splits der einzelnen Datensets der ReID-Auswertung

Datenset	Subset	IDs	Einzelbilder	Split	Kameras
Market1501	Training	751	12.936	40 %	6
	Query	750	3.368	10 %	6
	Galerie	751	15.913	50 %	6
OP0	Training	6	0	0 %	1
	Query	6	203	17 %	1
	Galerie	6	1.013	83 %	1
OP1	Training	6	0	0 %	1
	Query	6	200	17 %	1
	Galerie	6	999	83 %	1
OP1 Finetuning	Training	6	479	40 %	1
	Query	6	120	10 %	1
	Galerie	6	600	50 %	1
OP2	Training	6	0	0 %	1
	Query	6	210	17 %	1
	Galerie	6	1.048	83 %	1

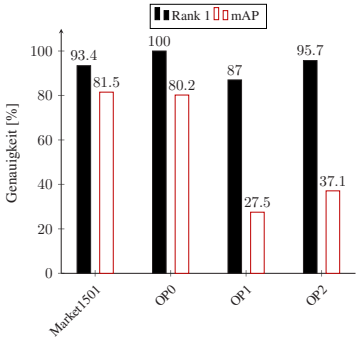


(a) OP0 **(b)** OP1 **(c)** OP2

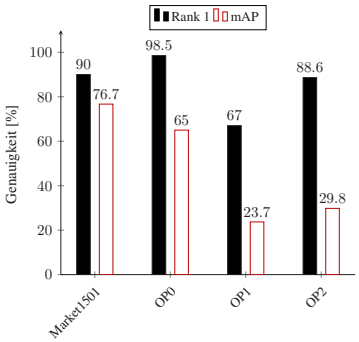
Abbildung A.11: Vergleich der getesteten Re-ID-Modelle bzgl. der einzelnen Datensets



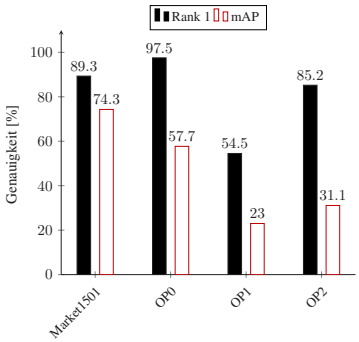
(a) OSNet



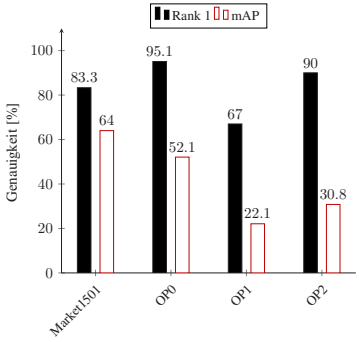
(b) OSNet-AIN



(c) PCB

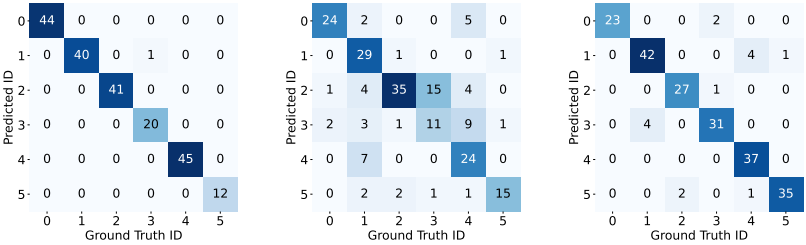


(d) ResNet-Mid



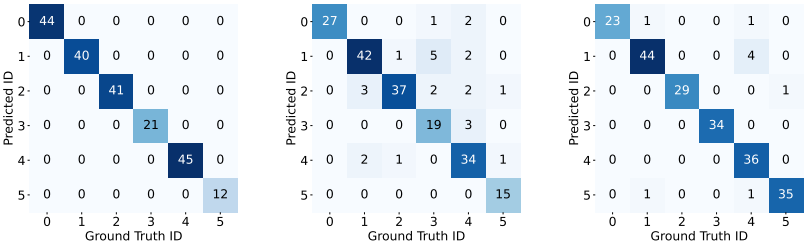
(e) MLFN

Abbildung A.12: Vergleich der untersuchten Datensets bzgl. der getesteten Re-ID-Modelle



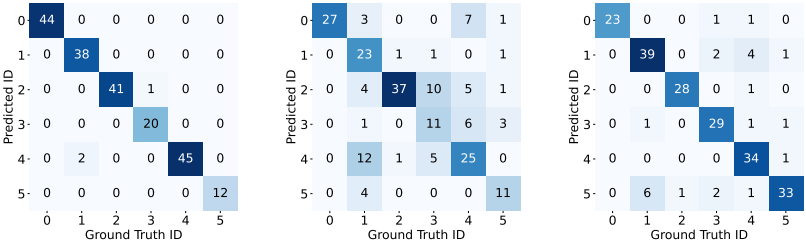
(a) OSNet OP0 **(b)** OSNet OP1 **(c)** OSNet OP2

Abbildung A.13: Konfusionsmatrizen des OSNet



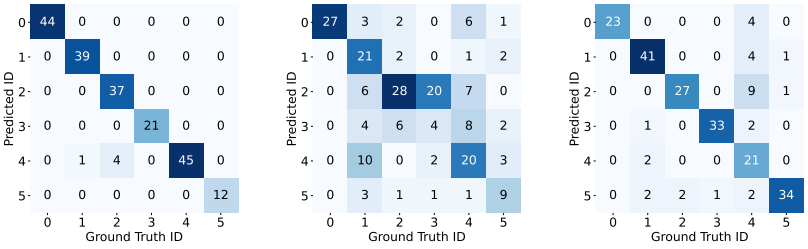
(a) OSNet-AIN OP0 **(b)** OSNet-AIN OP1 **(c)** OSNet-AIN OP2

Abbildung A.14: Konfusionsmatrizen des OSNet-AIN



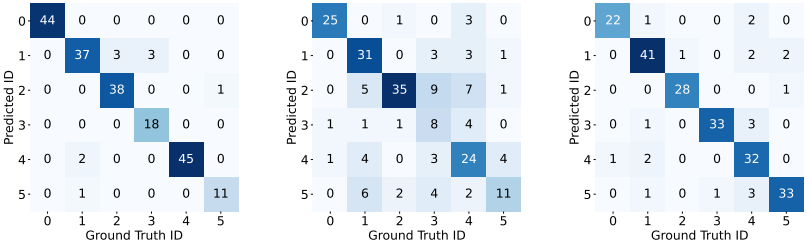
(a) PCB OP0 **(b)** PCB OP1 **(c)** PCB OP2

Abbildung A.15: Konfusionsmatrizen des PCB



(a) ResNet-Mid OP0 **(b)** ResNet-Mid OP1 **(c)** ResNet-Mid OP2

Abbildung A.16: Konfusionsmatrizen des ResNet-Mid



(a) MLFN OP0 **(b)** MLFN OP1 **(c)** MLFN OP2

Abbildung A.17: Konfusionsmatrizen des MLFN

A.3 Details zur Instrumentenerkennung

A.3.1 OP-Instrumente für die laparoskopische Cholezystektomie

Tabelle A.6: OP-Instrumente für die laparoskopische Cholezystektomie der Richard Wolf GmbH und deren Zuordnung für die Instrumenten- sowie Phasenerkennung dieser Arbeit.

Instrument	Klasse	HeiChole-Äquivalent	zugehörige Phase
Veress-Kanüle	Veress-Kanüle	-	2. Vorbereitung
Optik 10 mm, 30°	Optik	-	2. Vorbereitung
Optik 10 mm, 0°	Optik	-	2. Vorbereitung
Optik 5,3 mm, 30°	Optik	-	2. Vorbereitung
Fiber Lichtleiter	Fiber Lichtleiter	-	2. Vorbereitung
RIWO-ART-Trokarhülse 5,5 mm	Trokarhülse	-	2. Vorbereitung / 9. Abschluss
Trokarhülse 5,5 mm, selbsthaltend	Trokarhülse	-	2. Vorbereitung / 9. Abschluss
Trokarhülse 10 mm	Trokarhülse	-	2. Vorbereitung / 9. Abschluss
Trokar mit Schutzhülse	Trokar	-	2. Vorbereitung / 9. Abschluss
Trokar, kegelig-spitz	Trokar	-	2. Vorbereitung / 9. Abschluss
Dichtkonus 12-15 mm	Dichtkonus	-	9. Abschluss
Dilatations- & Führungshülse 5-10 mm	Hilfsinstrument	20 | Undefined instrument shaft	2. Vorbereitung / 8. Bergung Gallenblase
Instrumentierhülse & Extraktor 5-10 mm	Hilfsinstrument	20 | Undefined instrument shaft	2. Vorbereitung / 8. Bergung Gallenblase
Führungsstab 5 mm	Führungsstab	20 | Undefined instrument shaft	2. Vorbereitung / 8. Bergung Gallenblase
Greifzange mit Zahnreihe 5 mm	Modulare Instrumente	0 | Grasper	2. Vorbereitung - 9. Abschluss
Atraumatische Greifzange 5 mm	Modulare Instrumente	0 | Grasper	2. Vorbereitung - 9. Abschluss
Greifzange, stumpf, 5 mm	Modulare Instrumente	0 | Grasper	2. Vorbereitung - 9. Abschluss

Tabelle A.6: OP-Instrumente für die laparoskopische Cholezystektomie (Fortsetzung).

Instrument	Klasse	HeiChole-Äquivalent	zugehörige Phase
Greifzange, 2/3 krallig, 10 mm	Modulare Instrumente	0 | Grasper	2. Vorbereitung - 9. Abschluss
Atraumatische Tellerfasszange 10 mm	Modulare Instrumente	0 | Grasper	2. Vorbereitung - 9. Abschluss
Maryland Dissektor 5 mm	Modulare Instrumente	0 | Grasper	3. Dissekt. Calot Dreieck / 5. Dissekt. Gallenblase
Dissektionszange 10 mm	Modulare Instrumente	0 | Grasper	3. Dissekt. Calot Dreieck / 5. Dissekt. Gallenblase
Probe-Exzisionszange 5 mm	Modulare Instrumente	0 | Grasper	3. Dissekt. Calot Dreieck / 5. Dissekt. Gallenblase
Schere „Metzenbaum“ 5 mm	Modulare Instrumente	3 | Scissors	4. Clippen & Schneiden
Greif- & Präparierzange, bipolar	Modulare Instrumente	2 | Coagulation instruments	2. Vorbereitung - 9. Abschluss
Haken-Elektrode 5 mm	Koagulationsinstrumente	2 | Coagulation instruments	3. Dissekt. Calot Dreieck / 5. Dissekt. Gallenblase
HF-Anschlusskabel	HF-Anschlusskabel	-	3. Dissekt. Calot Dreieck / 5. Dissekt. Gallenblase
Clip-Applikator 10 mm	Clipper	1 | Clipper	4. Clippen & Schneiden
Ligaclips, VE=108 Stück	-	-	4. Clippen & Schneiden
Endo-Spreizer	Endo-Spreizer	-	8. Bergung Gallenblase
Saug-Spül-Rohr 5 mm	Saug-Spül-System	4 | Suction-irrigation	7. Blutstillung & Spülung
Saug-Spül-Handgriff	Saug-Spül-System	-	7. Blutstillung & Spülung
Verschlusskappe	Saug-Spül-System	-	7. Blutstillung & Spülung
Nadelzange	Nadelzange	-	9. Abschluss

Tabelle A.7: Verteilung der Klassen der OP-Instrumente für die laparoskopische Cholezystektomie jeweils in absoluter Häufigkeit und normalisiert über alle Klassen in Prozent. Die letzte Zeile gibt jeweils die Summe aller vorkommenden Instanzen im Teildatenset und den Anteil dieses Teildatensets am gesamten Datenset an.

Nr.	Klasse	Gesamt		Training		Validierung		Test	
		#	%	#	%	#	%	#	%
0	Endo-Spreizer	636	2,9	379	2,9	128	2,9	129	3,0
1	Modulare Instrumente	3.581	16,3	2.150	16,4	715	16,4	716	16,4
2	HF-Anschlusskabel	2.145	9,8	1.292	9,8	426	9,8	427	9,8
3	Trokarhülse	3.079	14,1	1.852	14,1	613	14,0	614	14,0
4	Koagulationsinstrument	1.609	7,3	966	7,4	321	7,4	322	7,4
5	Trokar	2.776	12,7	1.673	12,7	551	12,6	552	12,6
6	Clipper	684	3,1	408	3,1	138	3,2	138	3,2
7	Saug-Spül-System	1.351	6,1	810	6,2	271	6,2	270	6,2
8	Optik	1.035	4,7	619	4,7	208	4,8	208	4,8
9	Nadelzange	708	3,2	423	3,2	143	3,3	142	3,2
10	Fiber Lichtleiter	571	2,6	341	2,6	116	2,7	114	2,6
11	Führungsstab	750	3,4	453	3,4	148	3,4	149	3,4
12	Veress-Kanüle	737	3,4	447	3,4	145	3,3	145	3,3
13	Hilfsinstrument	1.483	6,8	898	6,8	292	6,7	293	6,7
14	Dichtkonus	737	3,4	448	3,4	144	3,3	145	3,3
15	Hintergrund	31	0,1	15	0,1	8	0,2	8	0,2
	Gesamt	21.913	100	13.174	60,1	4.367	19,9	4.372	20,0

A.3.2 Evaluation der Instrumentenerkennung

A.3.2.1 Konfusionsmatrizen

Die Labels der dargestellten Konfusionsmatrizen entsprechen der Nummerierung in Tabelle A.7.

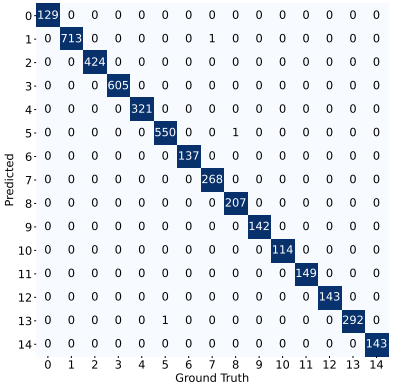


Abbildung A.18: Konfusionsmatrix YOLOv3

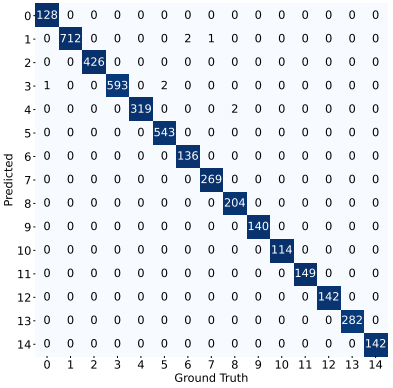


Abbildung A.19: Konfusionsmatrix YOLOv3-tiny

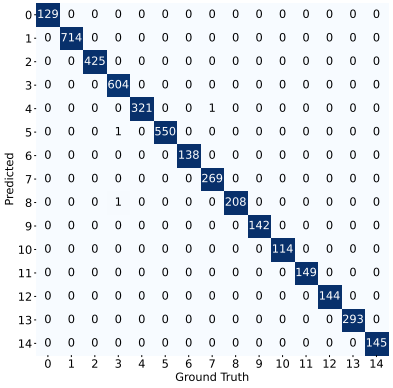


Abbildung A.20: Konfusionsmatrix YOLOv3-spp

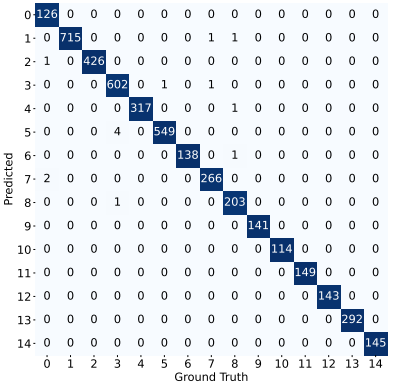


Abbildung A.21: Konfusionsmatrix YOLOv5n

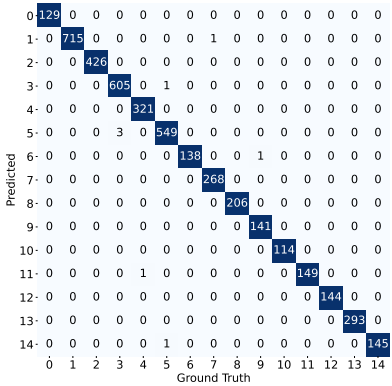


Abbildung A.22: Konfusionsmatrix YOLOv5s

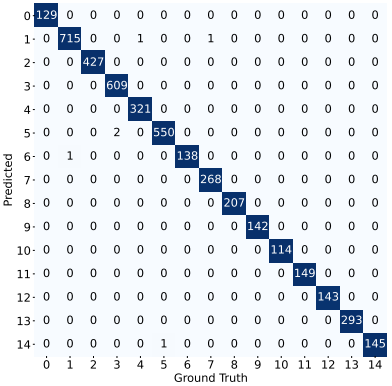


Abbildung A.23: Konfusionsmatrix YOLOv5l

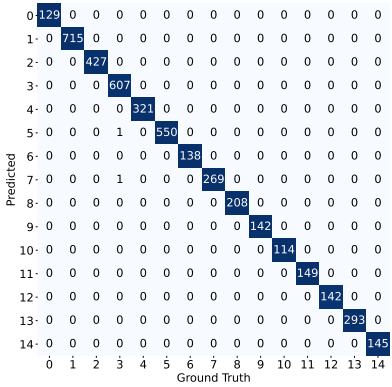


Abbildung A.24: Konfusionsmatrix YOLOv5x

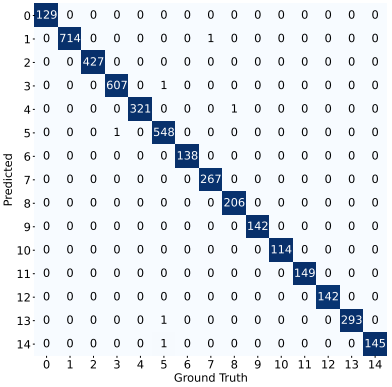


Abbildung A.25: Konfusionsmatrix YOLOv5n6 1280

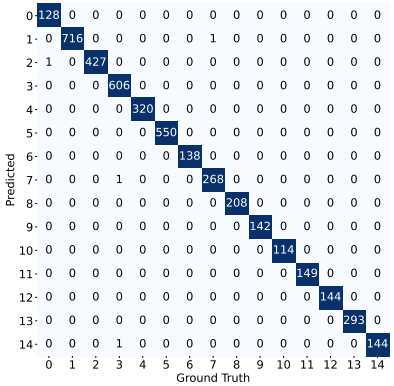


Abbildung A.26: Konfusionsmatrix YOLOv5s6 1280

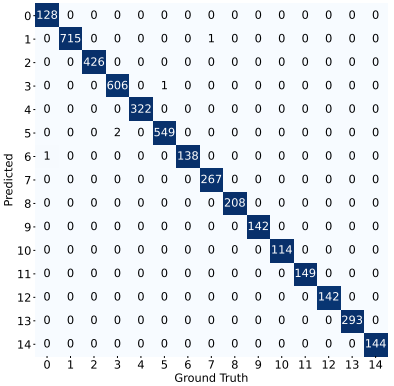


Abbildung A.27: Konfusionsmatrix YOLOv5m6 1280

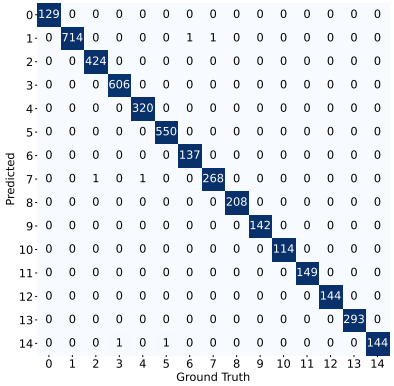


Abbildung A.28: Konfusionsmatrix YOLOv5l6 1280

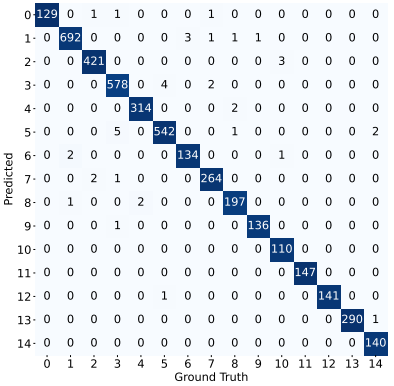


Abbildung A.29: Konfusionsmatrix YOLOv6n

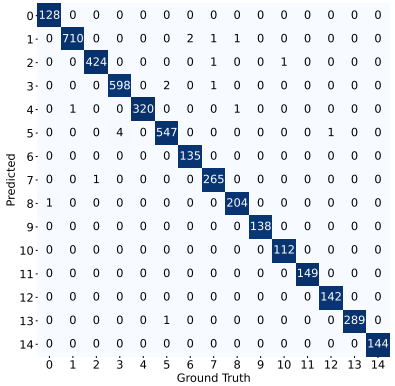


Abbildung A.30: Konfusionsmatrix YOLOv6s

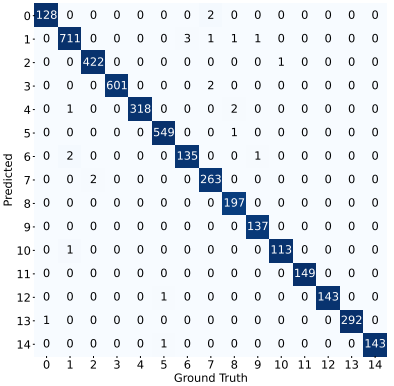


Abbildung A.31: Konfusionsmatrix YOLOv6m

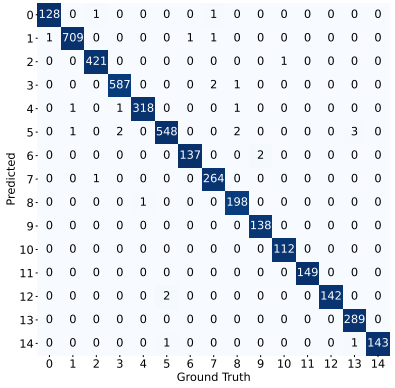


Abbildung A.32: Konfusionsmatrix YOLOv6l

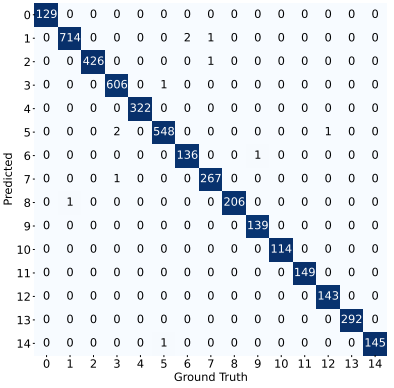


Abbildung A.33: Konfusionsmatrix YOLOv8n

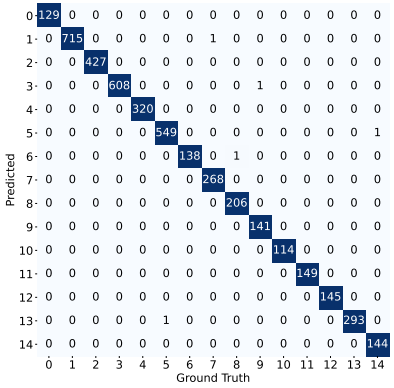


Abbildung A.34: Konfusionsmatrix YOLOv8s

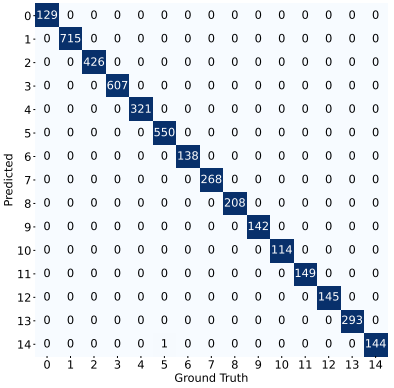


Abbildung A.35: Konfusionsmatrix YOLOv8m

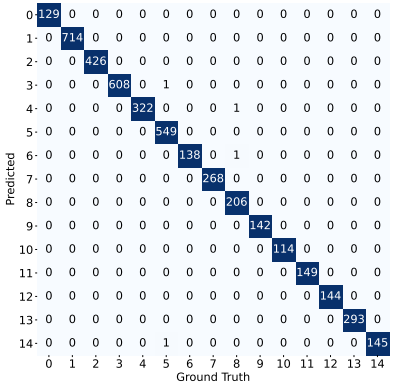


Abbildung A.36: Konfusionsmatrix YOLOv8l

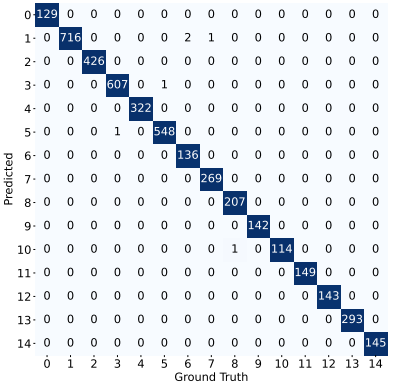


Abbildung A.37: Konfusionsmatrix YOLOv8x

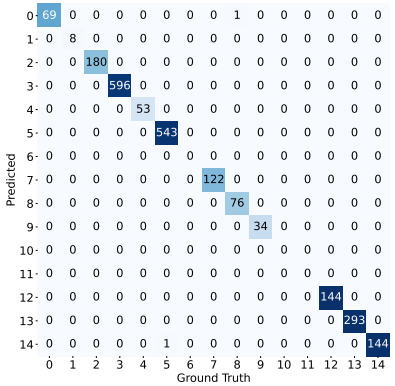


Abbildung A.38: Konfusionsmatrix RT-DETR-l

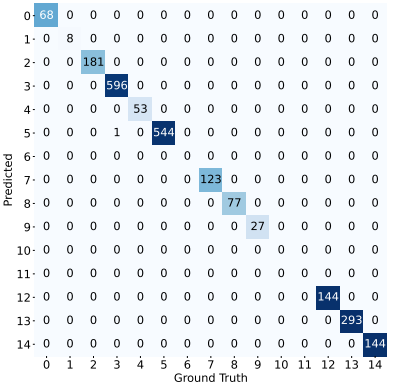


Abbildung A.39: Konfusionsmatrix RT-DETR-x

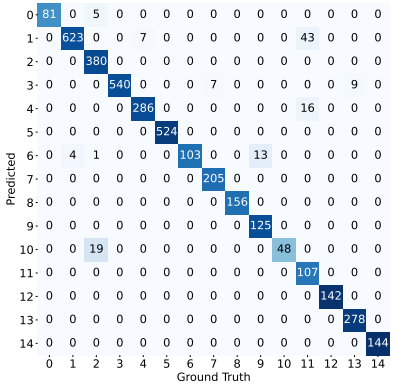


Abbildung A.40: Konfusionsmatrix YOLO-NAS-s

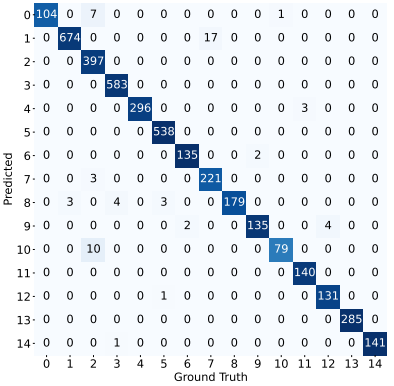


Abbildung A.41: Konfusionsmatrix YOLO-NAS-m

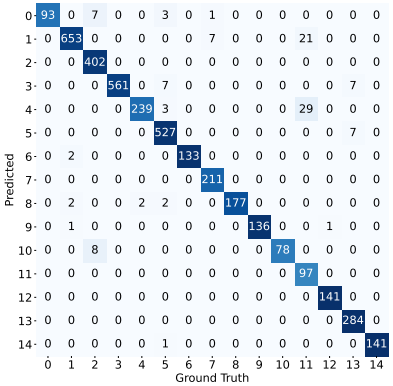


Abbildung A.42: Konfusionsmatrix YOLO-NAS-l

A.3.2.2 Precision-Recall-Kurven

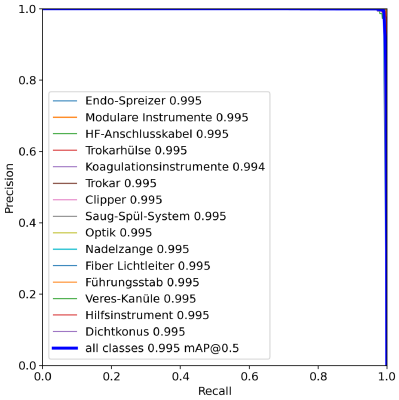


Abbildung A.43: Precision-Recall-Kurve YOLOv3

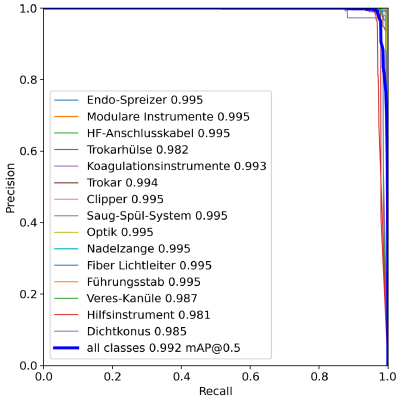


Abbildung A.44: Precision-Recall-Kurve YOLOv3-tiny

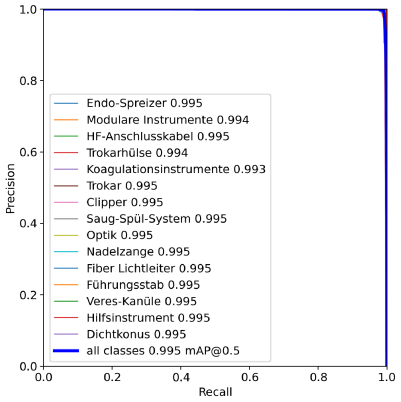


Abbildung A.45: Precision-Recall-Kurve YOLOv3spp

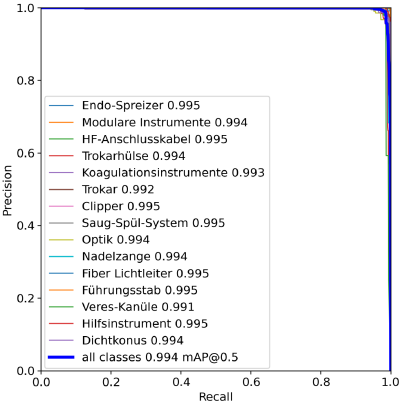


Abbildung A.46: Precision-Recall-Kurve YOLOv5n

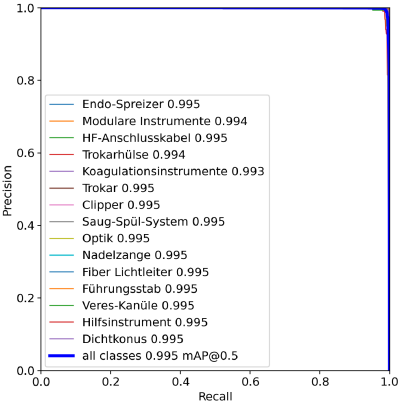


Abbildung A.47: Precision-Recall-Kurve YOLOv5s

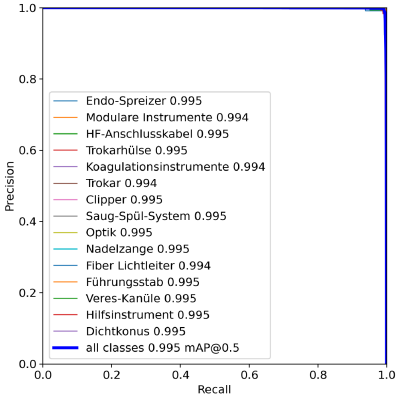


Abbildung A.48: Precision-Recall-Kurve YOLOv5l

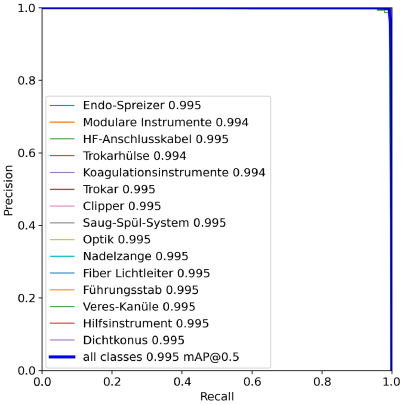


Abbildung A.49: Precision-Recall-Kurve YOLOv5x

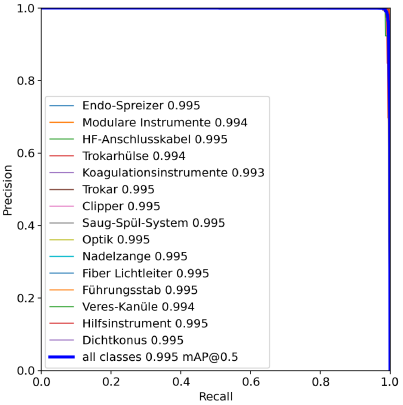


Abbildung A.50: Precision-Recall-Kurve YOLOv5n 1280

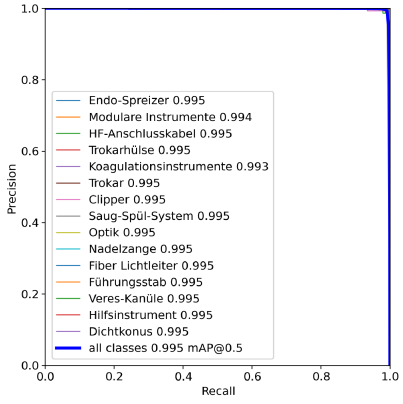


Abbildung A.51: Precision-Recall-Kurve YOLOv5m 1280

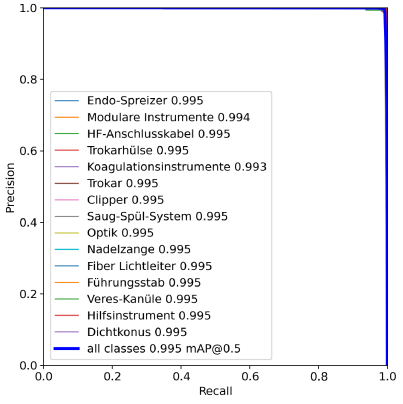


Abbildung A.52: Precision-Recall-Kurve YOLOv5l 1280

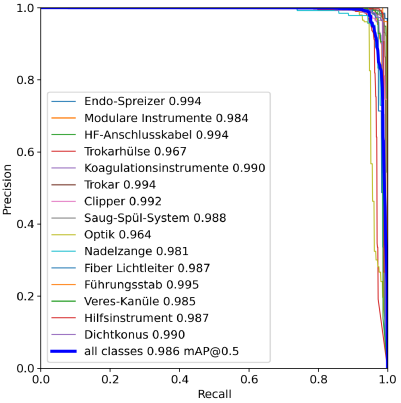


Abbildung A.53: Precision-Recall-Kurve YOLOv6n

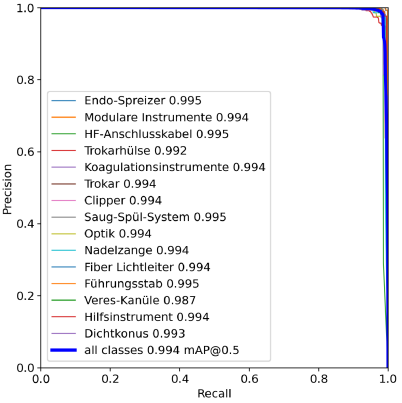


Abbildung A.54: Precision-Recall-Kurve YOLOv6s

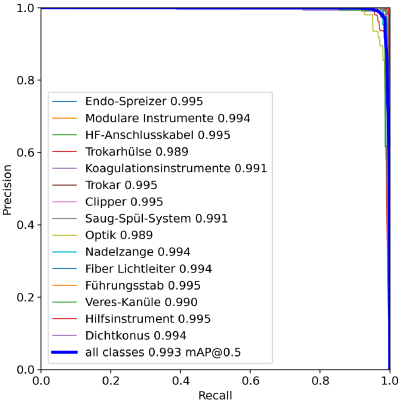


Abbildung A.55: Precision-Recall-Kurve YOLOv6m

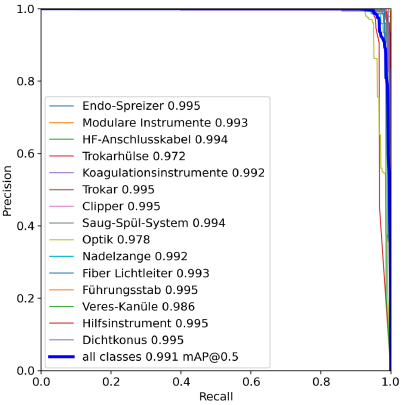


Abbildung A.56: Precision-Recall-Kurve YOLOv6l

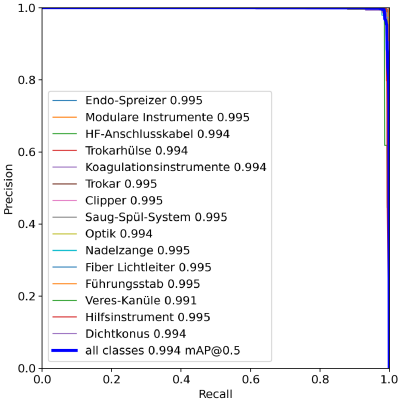


Abbildung A.57: Precision-Recall-Kurve YOLOv8n

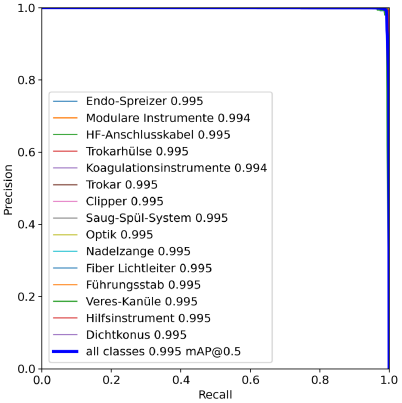


Abbildung A.58: Precision-Recall-Kurve YOLOv8s

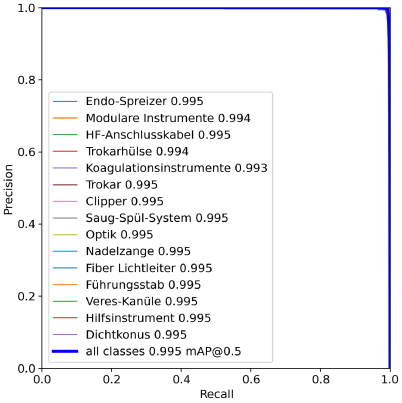


Abbildung A.59: Precision-Recall-Kurve YOLOv8x

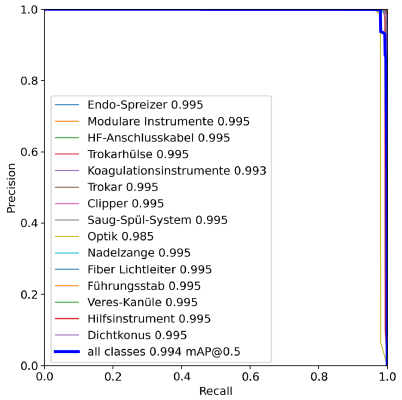


Abbildung A.60: Precision-Recall-Kurve RT-DETR-l

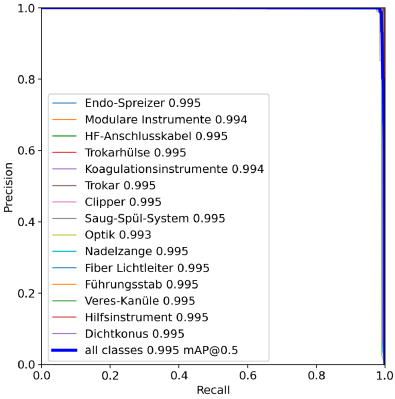


Abbildung A.61: Precision-Recall-Kurve
RT-DETR-x

A.3.2.3 Precision-Kurven

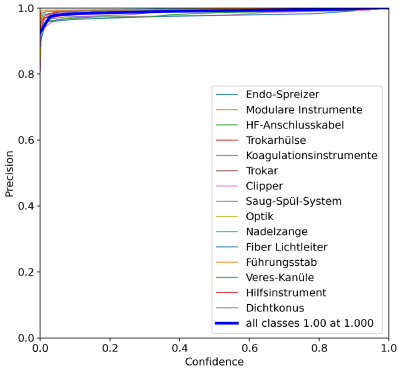


Abbildung A.62: Precision-Kurve
YOLOv3

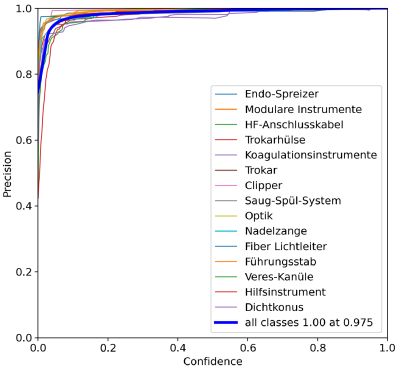


Abbildung A.63: Precision-Kurve
YOLOv3-tiny

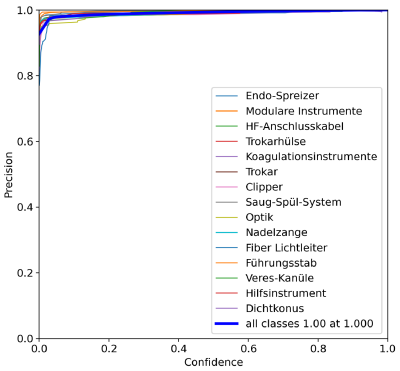


Abbildung A.64: Precision-Kurve YOLOv3spp

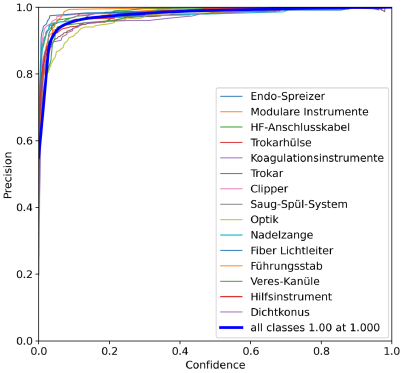


Abbildung A.65: Precision-Kurve YOLOv5n

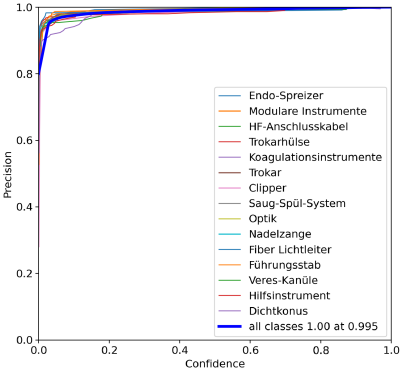


Abbildung A.66: Precision-Kurve YOLOv5s

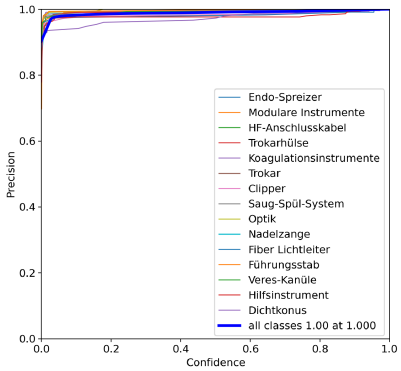


Abbildung A.67: Precision-Kurve YOLOv5l

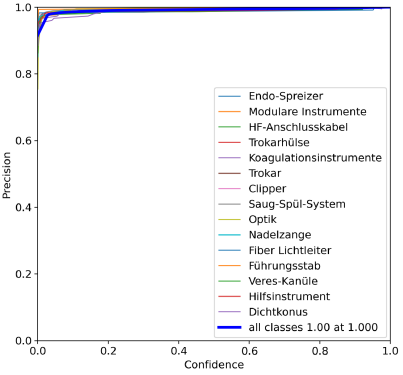


Abbildung A.68: Precision-Kurve YOLOv5x

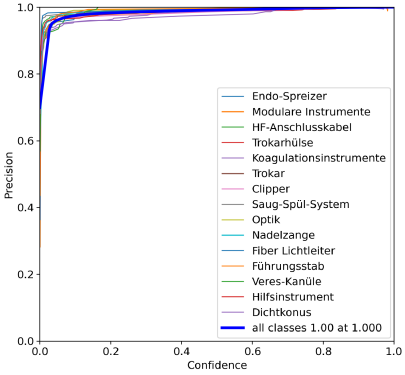


Abbildung A.69: Precision-Kurve YOLOv5n 1280

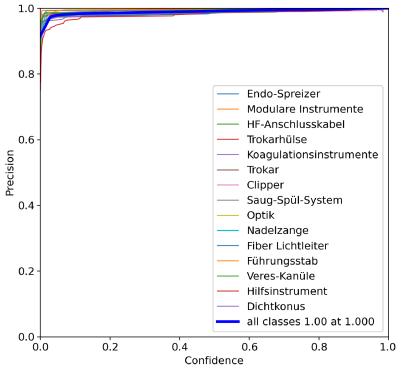


Abbildung A.70: Precision-Kurve YOLOv5m 1280

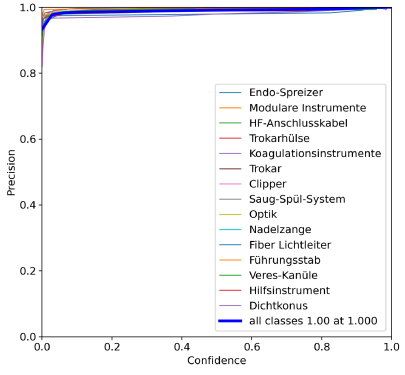


Abbildung A.71: Precision-Kurve YOLOv5l 1280

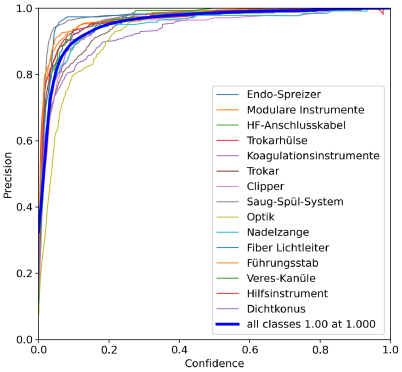


Abbildung A.72: Precision-Kurve YOLOv6n

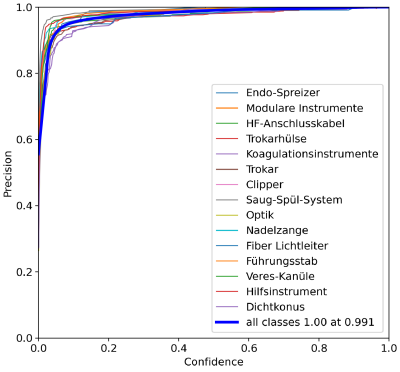


Abbildung A.73: Precision-Kurve YOLOv6s

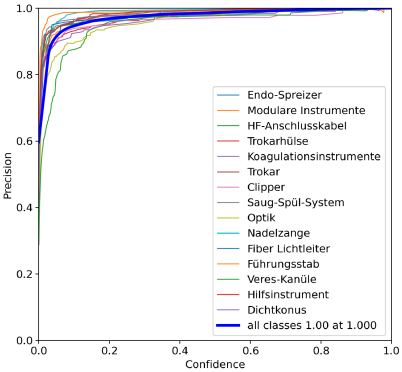


Abbildung A.74: Precision-Kurve YOLOv6m

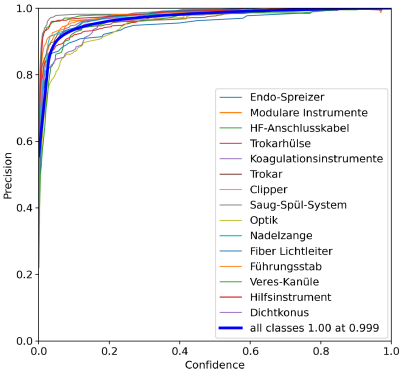


Abbildung A.75: Precision-Kurve YOLOv6l

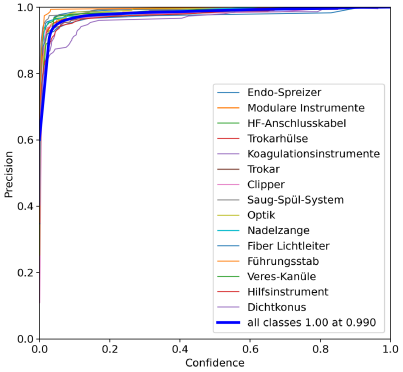


Abbildung A.76: Precision-Kurve YOLOv8n

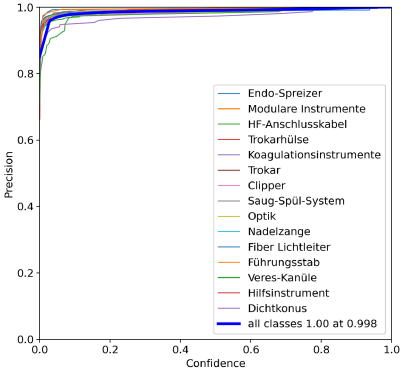


Abbildung A.77: Precision-Kurve YOLOv8s

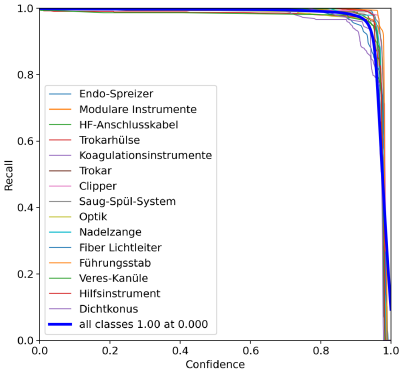


Abbildung A.78: Precision-Kurve YOLOv8x

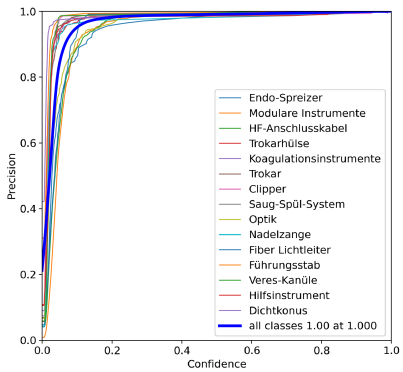


Abbildung A.79: Precision-Kurve RT-DETR-l

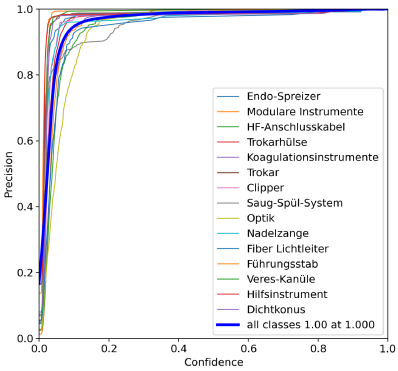


Abbildung A.80: Precision-Kurve RT-DETR-x

A.3.2.4 Recall-Kurven

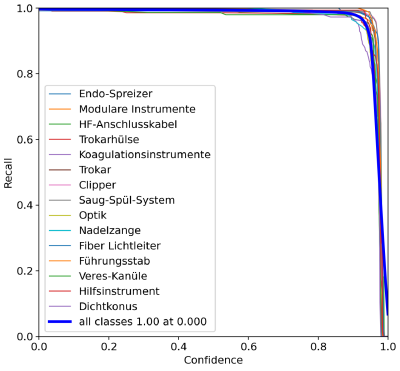


Abbildung A.81: Recall-Kurve YOLOv3

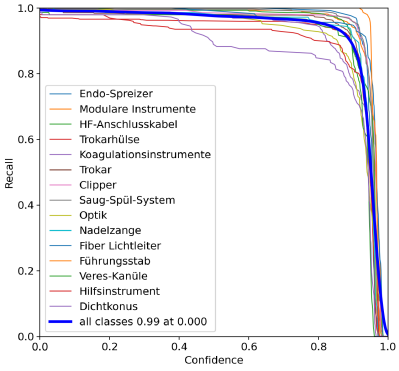


Abbildung A.82: Recall-Kurve YOLOv3-tiny

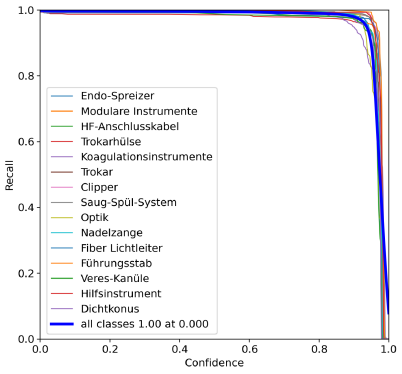


Abbildung A.83: Recall-Kurve YOLOv3spp

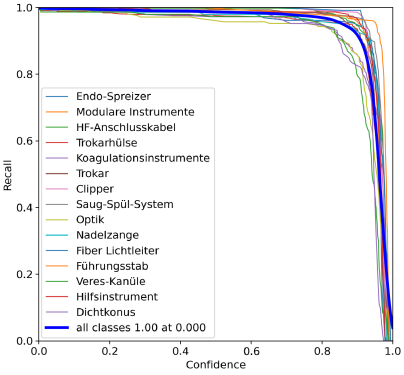


Abbildung A.84: Recall-Kurve YOLOv5n

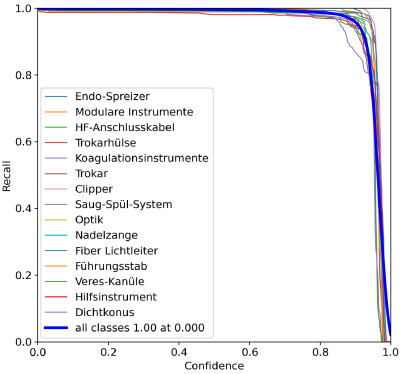


Abbildung A.85: Recall-Kurve YOLOv5s

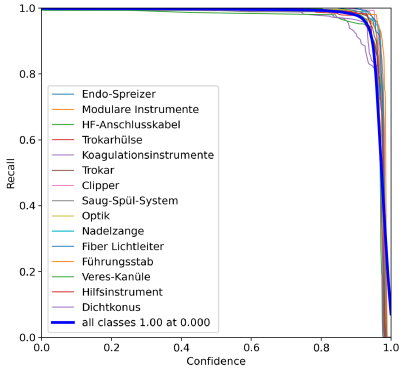


Abbildung A.86: Recall-Kurve YOLOv5l

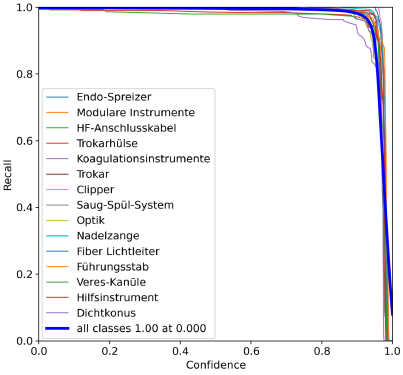


Abbildung A.87: Recall-Kurve YOLOv5x

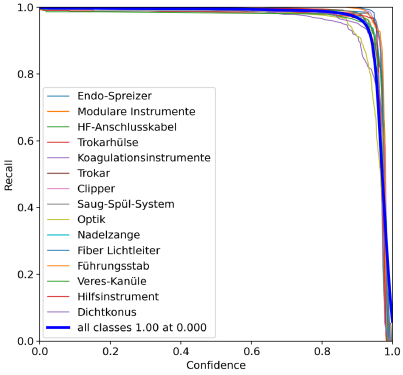


Abbildung A.88: Recall-Kurve YOLOv5n 1280

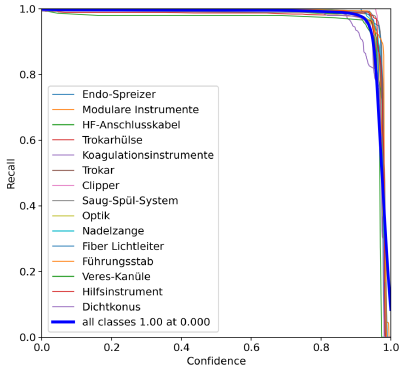


Abbildung A.89: Recall-Kurve YOLOv5m 1280

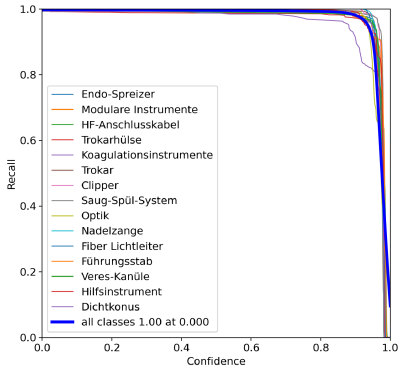


Abbildung A.90: Recall-Kurve YOLOv5l 1280

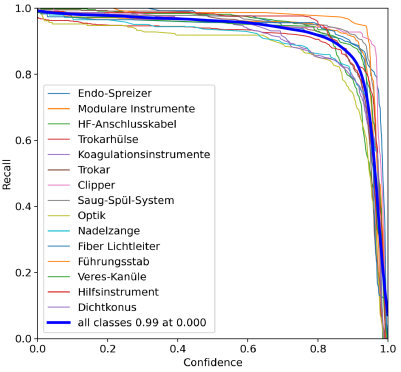


Abbildung A.91: Recall-Kurve YOLOv6n

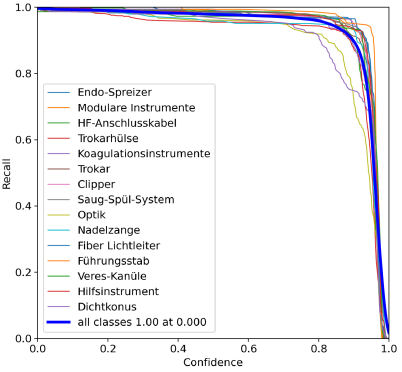


Abbildung A.92: Recall-Kurve YOLOv6s

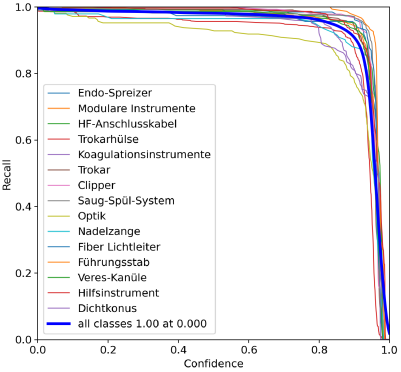


Abbildung A.93: Recall-Kurve YOLOv6m

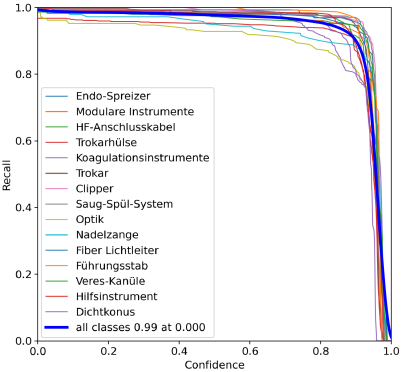


Abbildung A.94: Recall-Kurve YOLOv6l

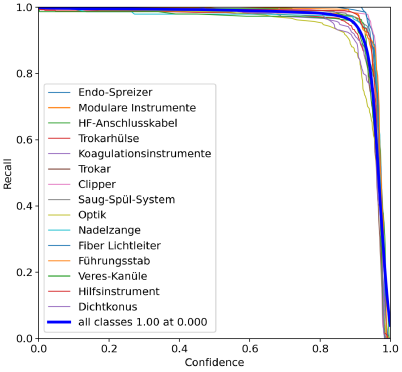


Abbildung A.95: Recall-Kurve YOLOv8n

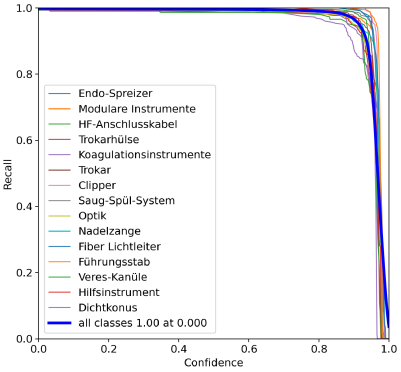


Abbildung A.96: Recall-Kurve YOLOv8s

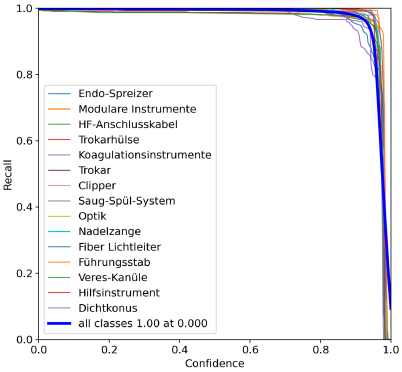


Abbildung A.97: Recall-Kurve YOLOv8x

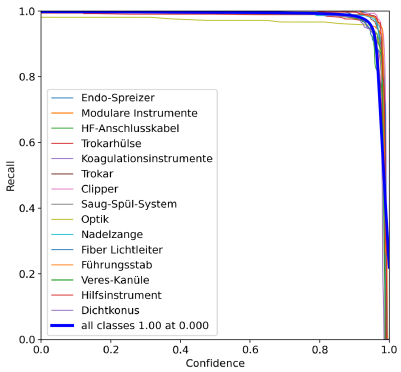


Abbildung A.98: Recall-Kurve RT-DETR-l

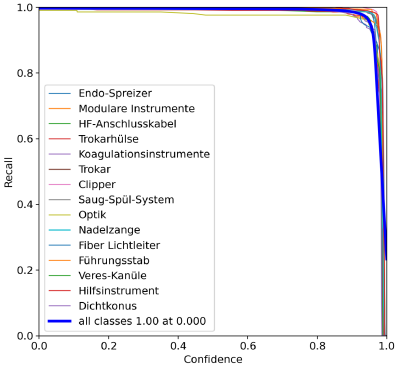


Abbildung A.99: Recall-Kurve RT-DETR-x

A.3.2.5 F1-Kurven

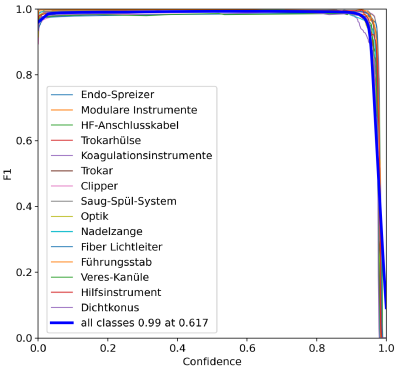


Abbildung A.100: F1-Kurve YOLOv3

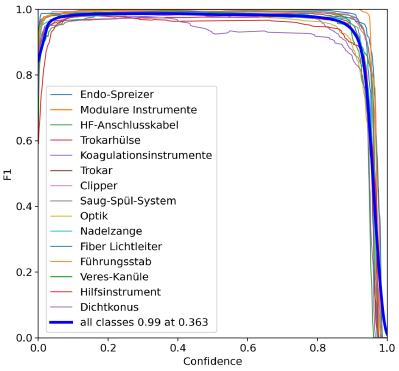


Abbildung A.101: F1-Kurve YOLOv3-tiny

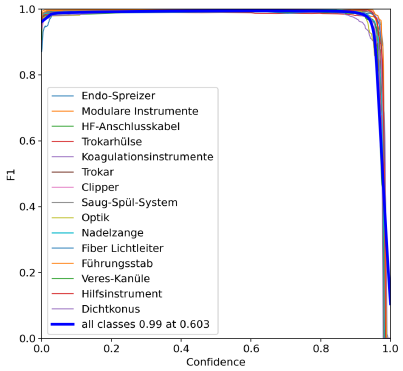


Abbildung A.102: F1-Kurve YOLOv3spp

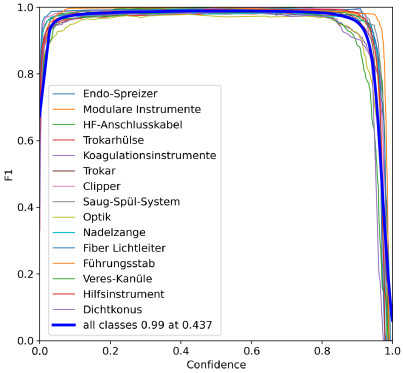


Abbildung A.103: F1-Kurve YOLOv5n

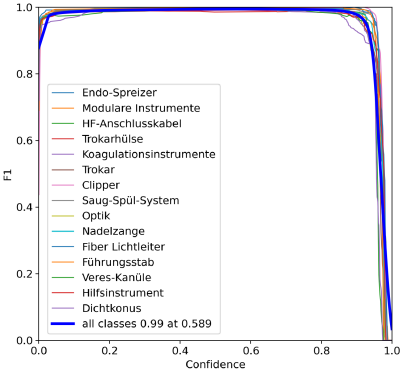


Abbildung A.104: F1-Kurve YOLOv5s

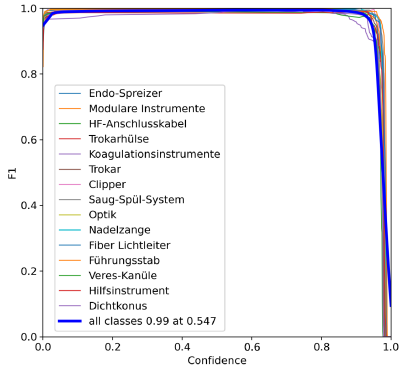


Abbildung A.105: F1-Kurve YOLOv5l

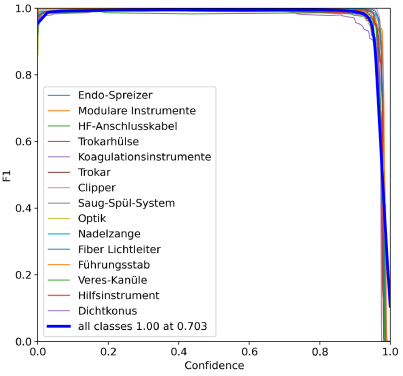


Abbildung A.106: F1-Kurve YOLOv5x

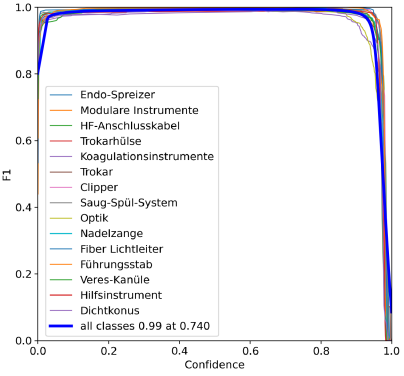


Abbildung A.107: F1-Kurve YOLOv5n 1280

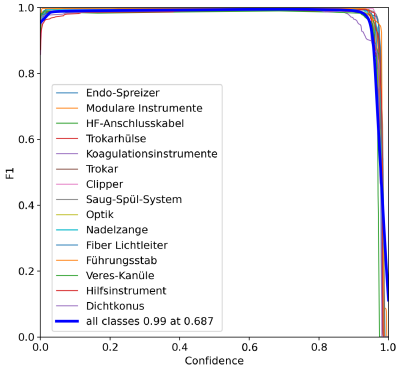


Abbildung A.108: F1-Kurve YOLOv5m 1280

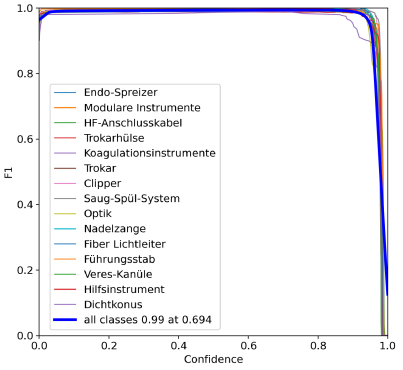


Abbildung A.109: F1-Kurve YOLOv5l 1280

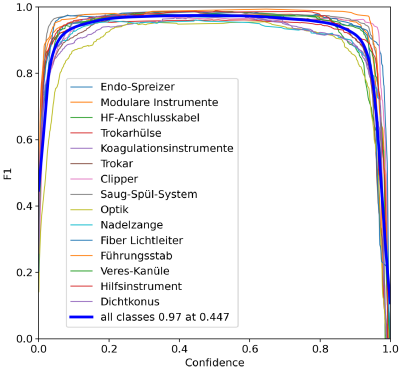


Abbildung A.110: F1-Kurve YOLOv6n

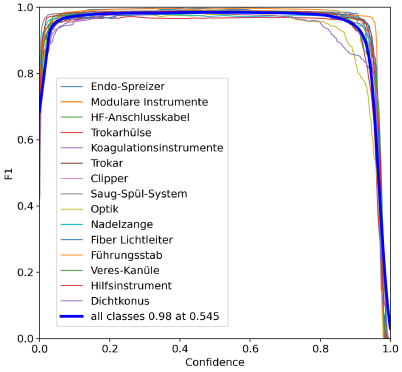


Abbildung A.111: F1-Kurve YOLOv6s

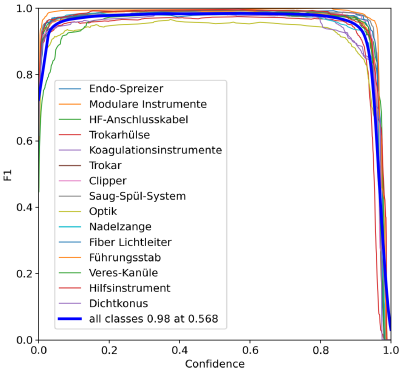


Abbildung A.112: F1-Kurve YOLOv6m

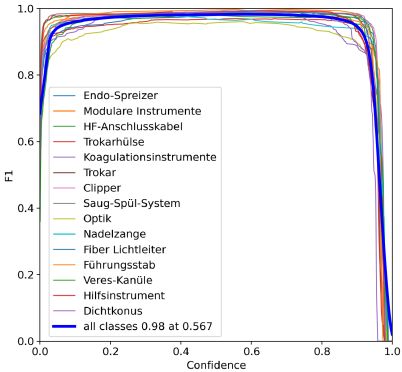


Abbildung A.113: F1-Kurve YOLOv6l

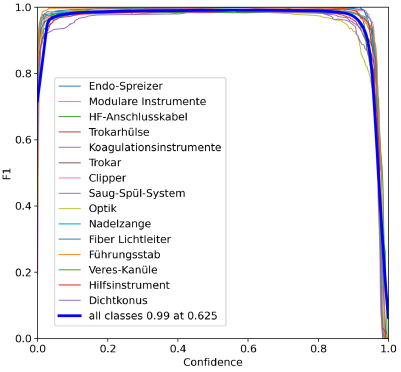


Abbildung A.114: F1-Kurve YOLOv8n

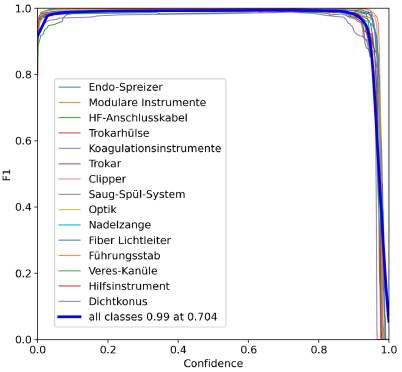


Abbildung A.115: F1-Kurve YOLOv8s

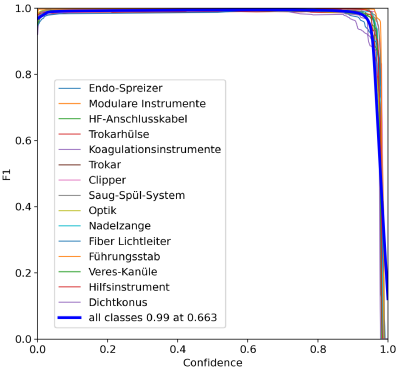


Abbildung A.116: F1-Kurve YOLOv8x

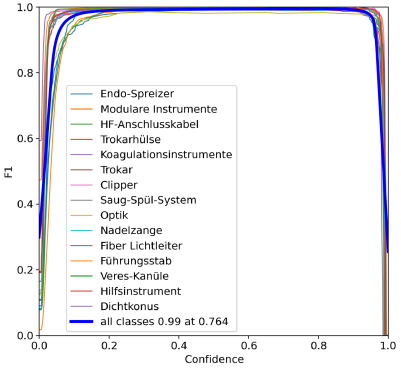


Abbildung A.117: F1-Kurve RT-DETR-l

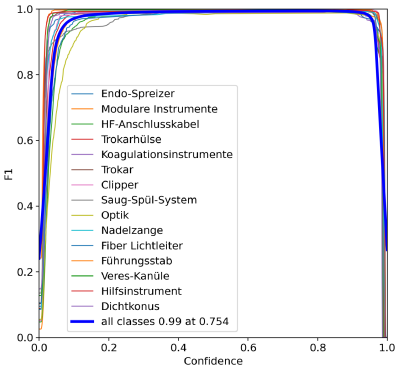


Abbildung A.118: F1-Kurve RT-DETR-x

A.4 Evaluation Endoskopauswertung

A.4.1 Phasen-Zeit-Diagramme Cholec80

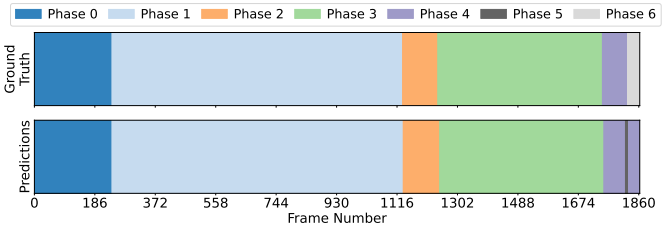


Abbildung A.119: Phasen-Zeit-Diagramm Cholec80-Video 65

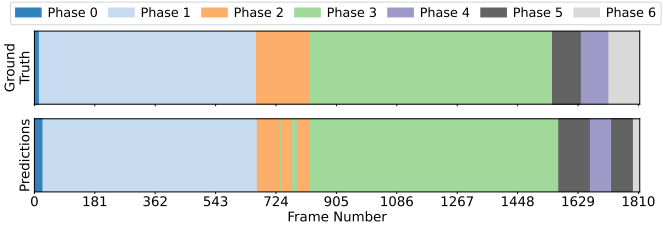


Abbildung A.120: Phasen-Zeit-Diagramm Cholec80-Video 66

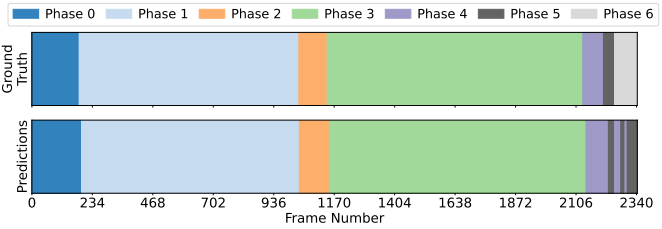


Abbildung A.121: Phasen-Zeit-Diagramm Cholec80-Video 67

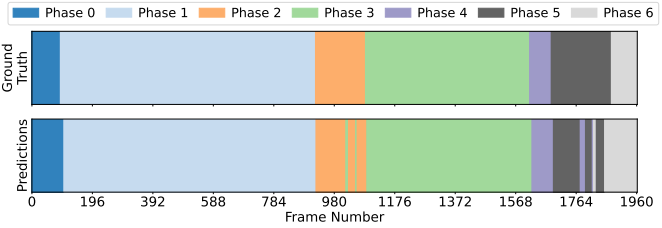


Abbildung A.122: Phasen-Zeit-Diagramm Cholec80-Video 68

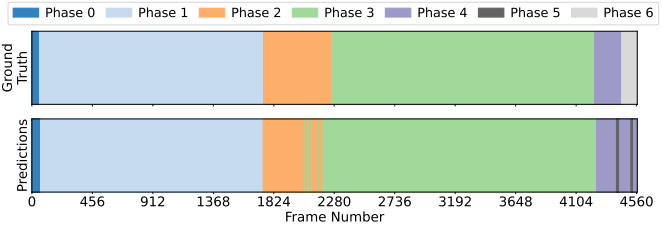


Abbildung A.123: Phasen-Zeit-Diagramm Cholec80-Video 69

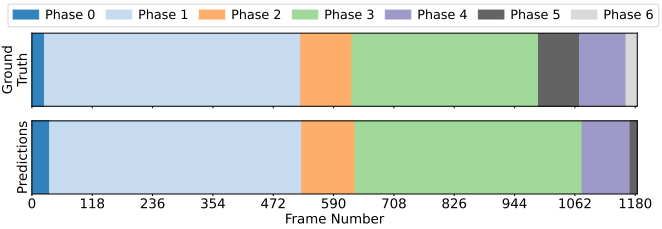


Abbildung A.124: Phasen-Zeit-Diagramm Cholec80-Video 70

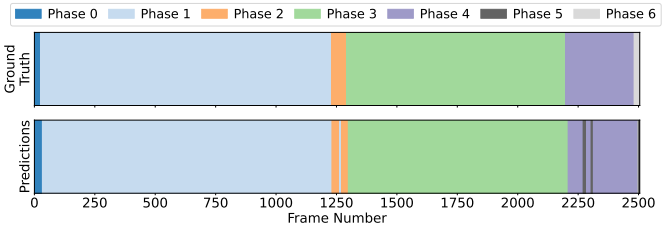


Abbildung A.125: Phasen-Zeit-Diagramm Cholec80-Video 71

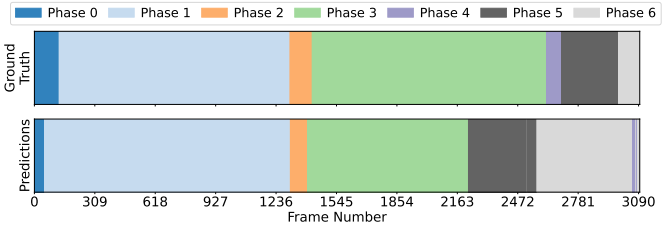


Abbildung A.126: Phasen-Zeit-Diagramm Cholec80-Video 72

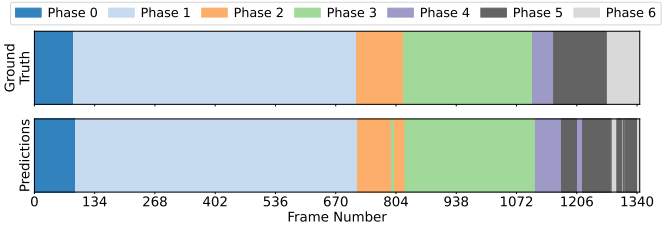


Abbildung A.127: Phasen-Zeit-Diagramm Cholec80-Video 73

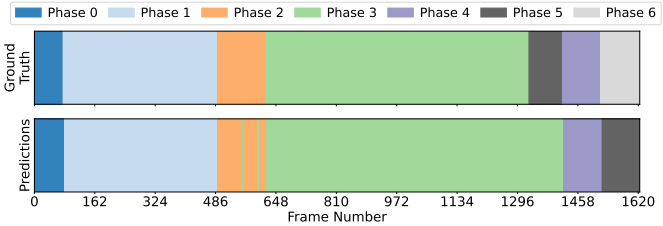


Abbildung A.128: Phasen-Zeit-Diagramm Cholec80-Video 74

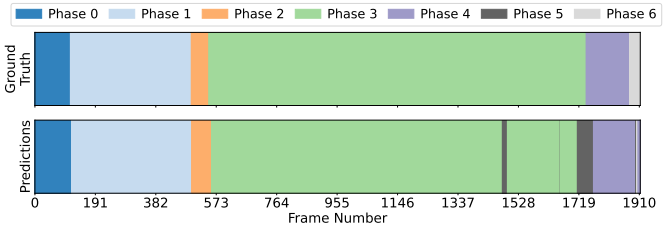


Abbildung A.129: Phasen-Zeit-Diagramm Cholec80-Video 75

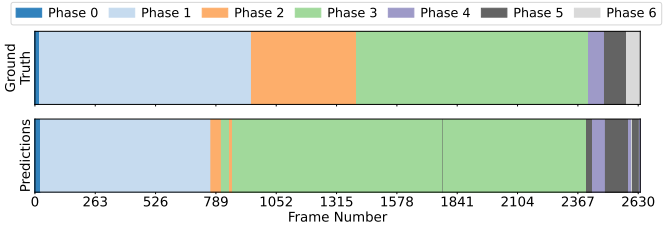


Abbildung A.130: Phasen-Zeit-Diagramm Cholec80-Video 76

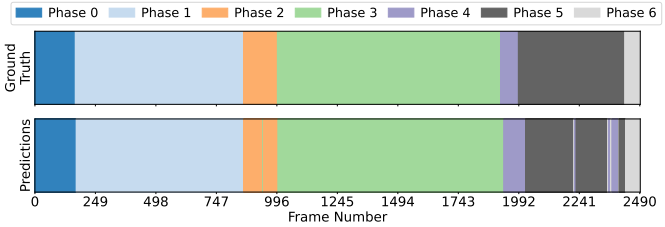


Abbildung A.131: Phasen-Zeit-Diagramm Cholec80-Video 77

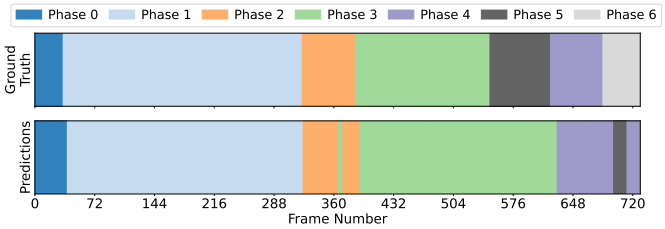


Abbildung A.132: Phasen-Zeit-Diagramm Cholec80-Video 78

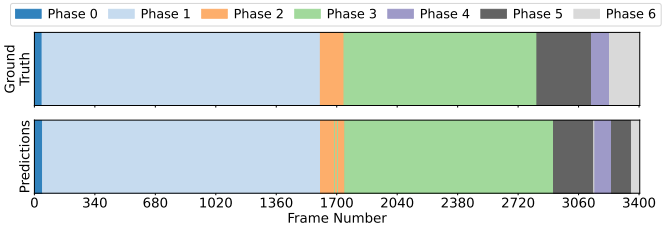


Abbildung A.133: Phasen-Zeit-Diagramm Cholec80-Video 79

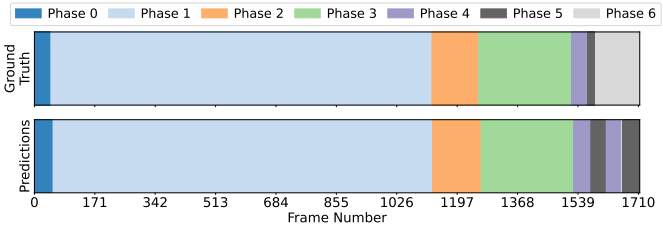


Abbildung A.134: Phasen-Zeit-Diagramm Cholec80-Video 80

A.4.2 Phasen-Zeit-Diagramme HeiCHole

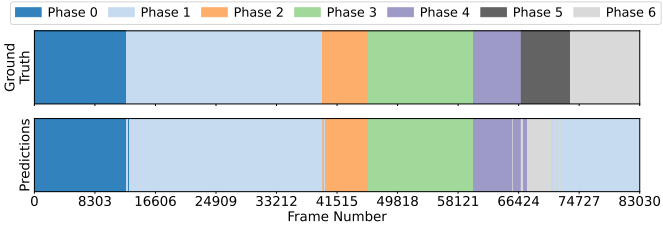


Abbildung A.135: Phasen-Zeit-Diagramm HeiChole-Video 20

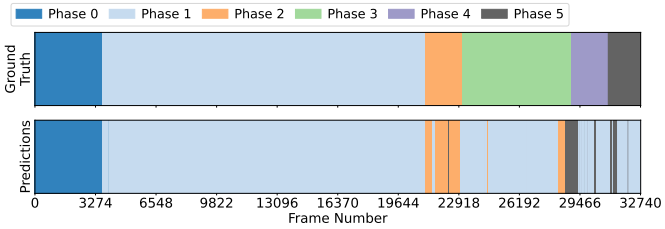


Abbildung A.136: Phasen-Zeit-Diagramm HeiChole-Video 21

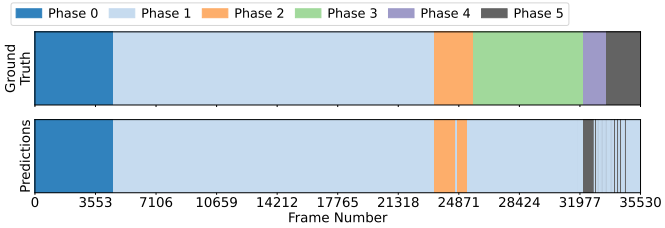


Abbildung A.137: Phasen-Zeit-Diagramm HeiChole-Video 22

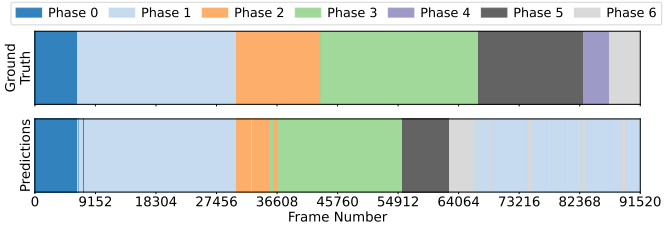


Abbildung A.138: Phasen-Zeit-Diagramm HeiChole-Video 23

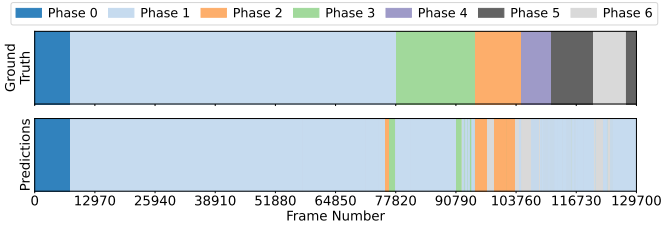


Abbildung A.139: Phasen-Zeit-Diagramm HeiChole-Video 24

A.4.3 Phasen-Zeit-Diagramme Cholec80 + HeiChole

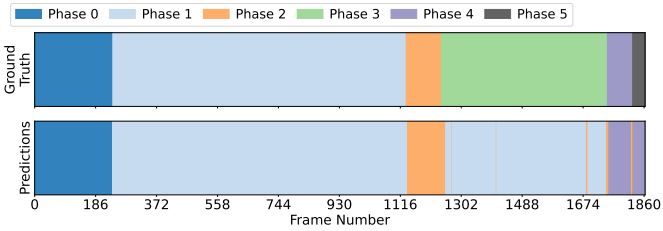


Abbildung A.140: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 65

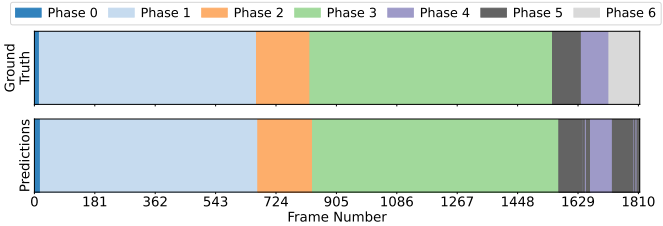


Abbildung A.141: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 66

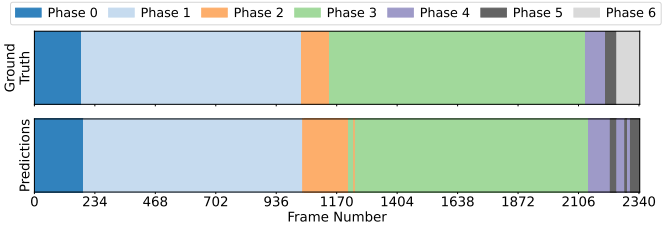


Abbildung A.142: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 67

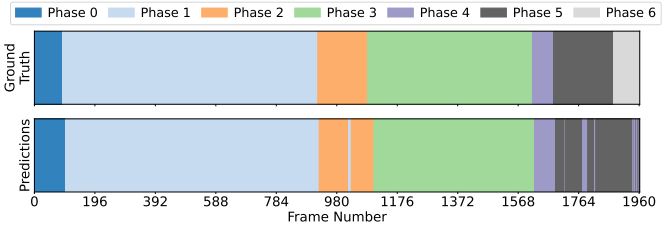


Abbildung A.143: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 68

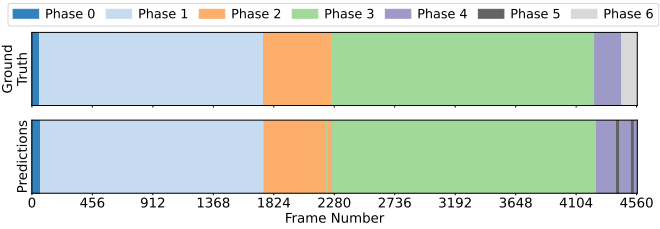


Abbildung A.144: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 69

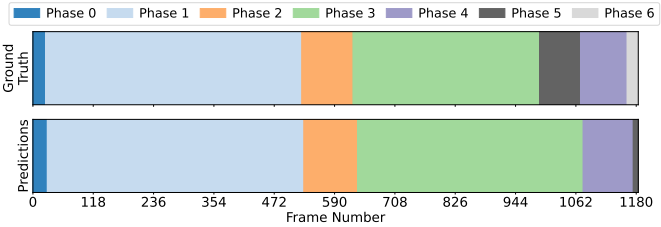


Abbildung A.145: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 70

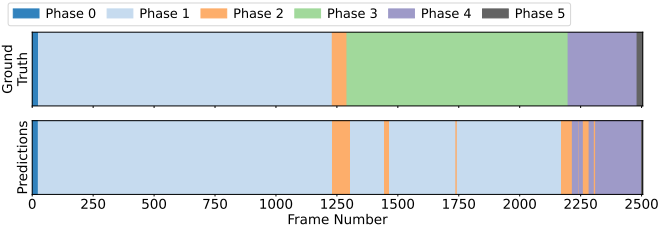


Abbildung A.146: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 71

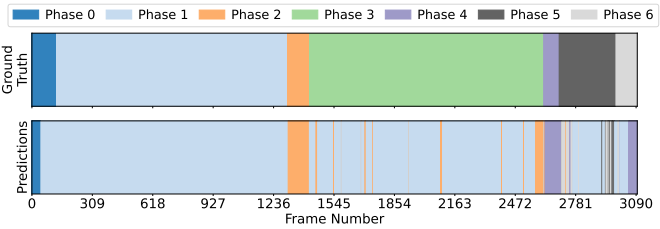


Abbildung A.147: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 72

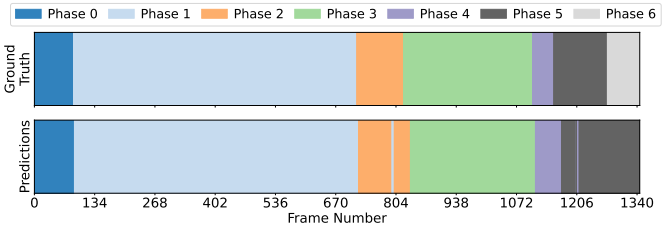


Abbildung A.148: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 73

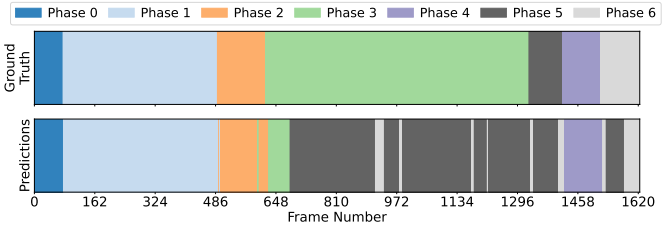


Abbildung A.149: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 74

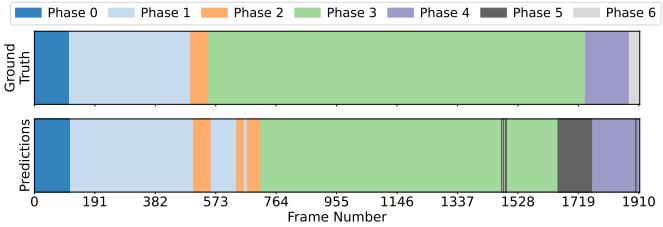


Abbildung A.150: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 75

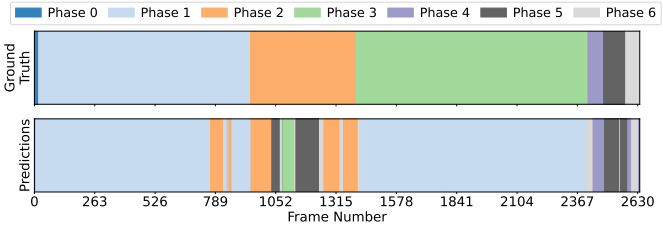


Abbildung A.151: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 76

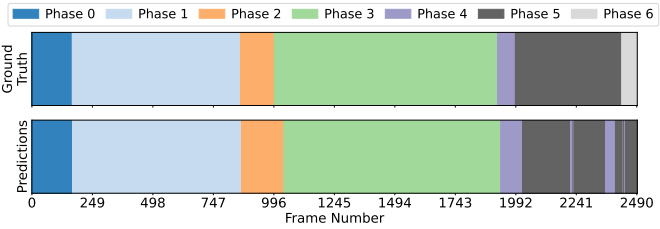


Abbildung A.152: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 77

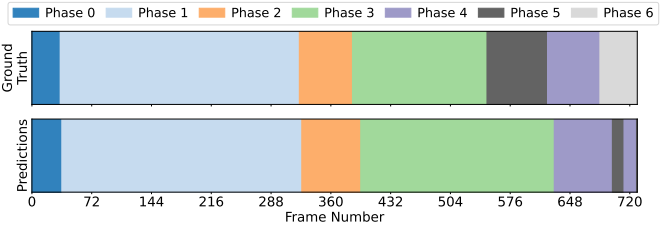


Abbildung A.153: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 78

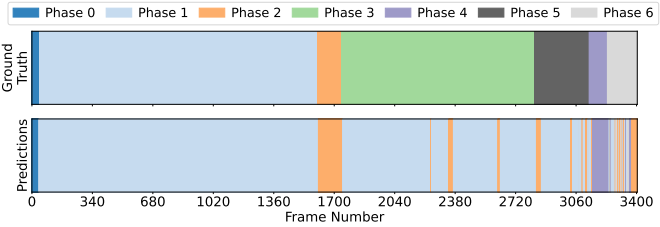


Abbildung A.154: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 79

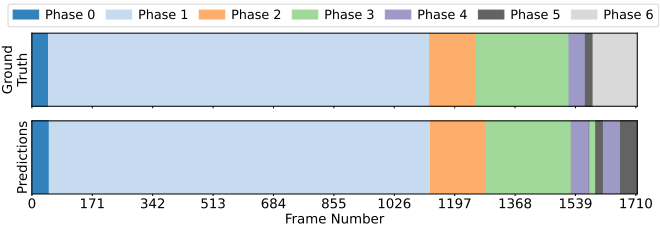


Abbildung A.155: Phasen-Zeit-Diagramm Cholec80 + HeiChole, Cholec80-Video 80

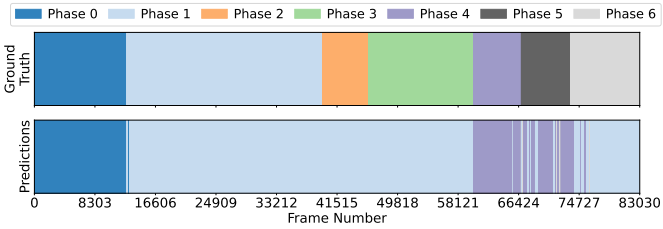


Abbildung A.156: Phasen-Zeit-Diagramm Cholec80 + HeiChole, HeiChole-Video 20

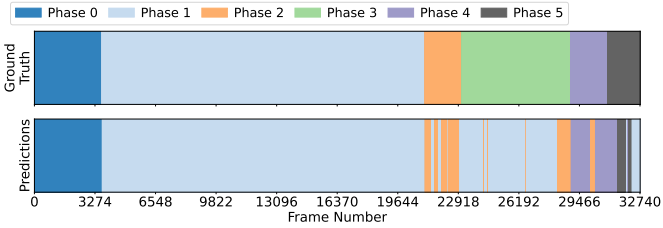


Abbildung A.157: Phasen-Zeit-Diagramm Cholec80 + HeiChole, HeiChole-Video 21

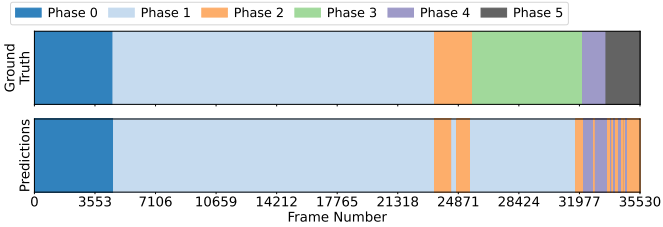


Abbildung A.158: Phasen-Zeit-Diagramm Cholec80 + HeiChole-Video, HeiChole-Video 22

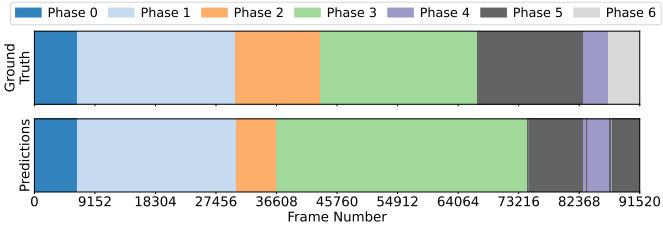


Abbildung A.159: Phasen-Zeit-Diagramm Cholec80 + HeiChole-Video, HeiChole-Video 23

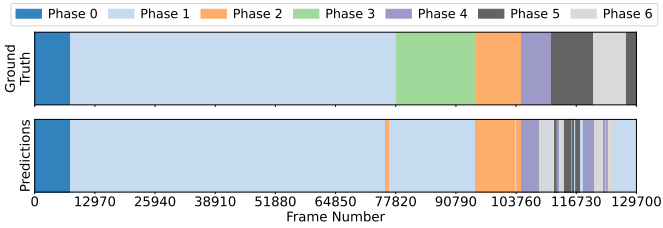


Abbildung A.160: Phasen-Zeit-Diagramm Cholec80 + HeiChole, HeiChole-Video 24

A.4.4 Phasen-Zeit-Diagramme Cholec80 + Finetuning auf HeiChole

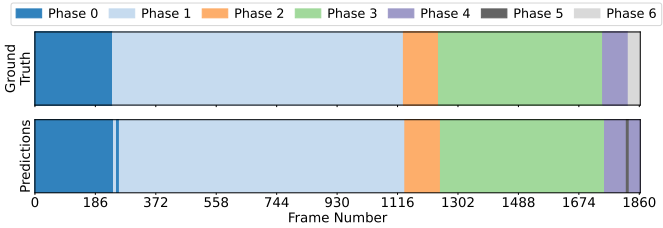


Abbildung A.161: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 65

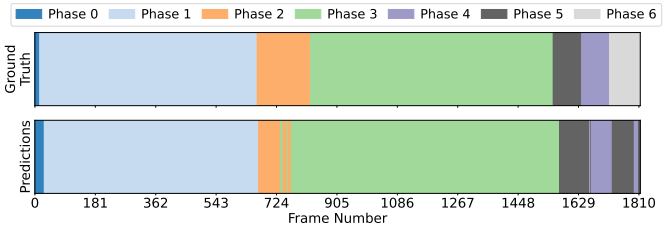


Abbildung A.162: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 66

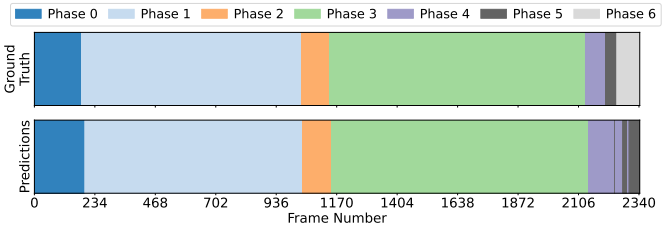


Abbildung A.163: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 67

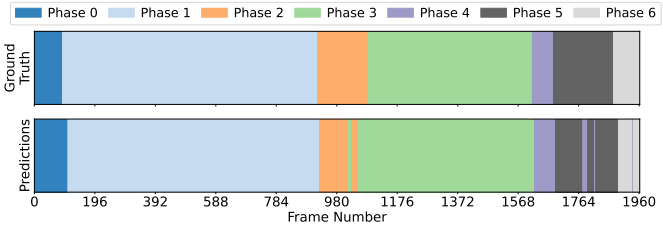


Abbildung A.164: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 68

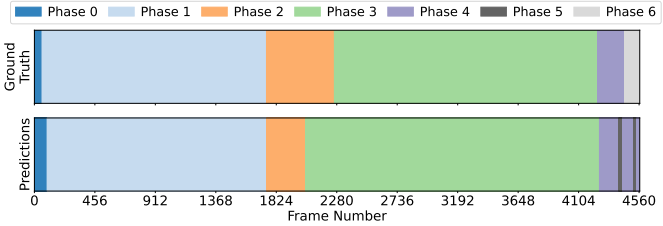


Abbildung A.165: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 69

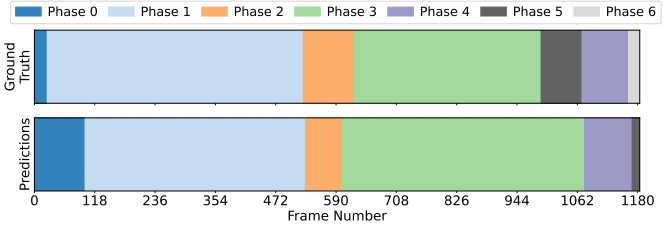


Abbildung A.166: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 70

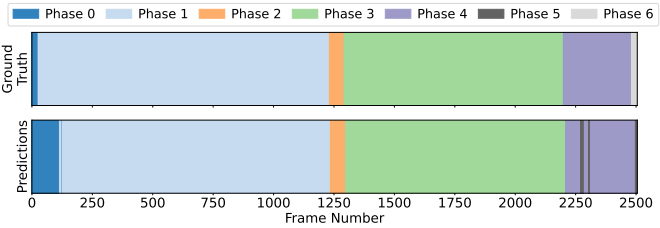


Abbildung A.167: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 71

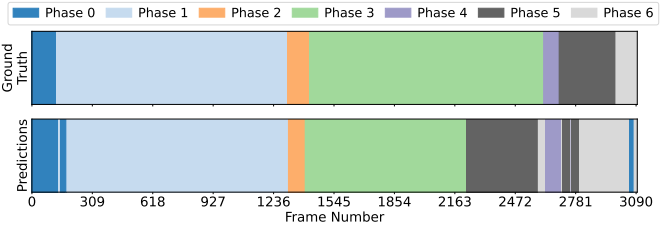


Abbildung A.168: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 72

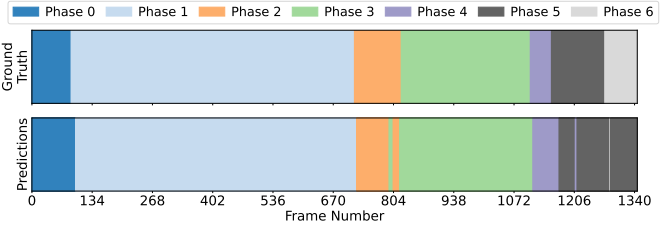


Abbildung A.169: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 73

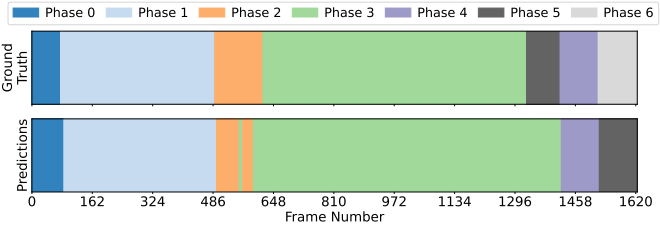


Abbildung A.170: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 74

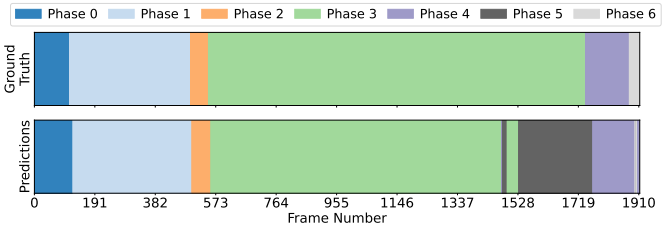


Abbildung A.171: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 75

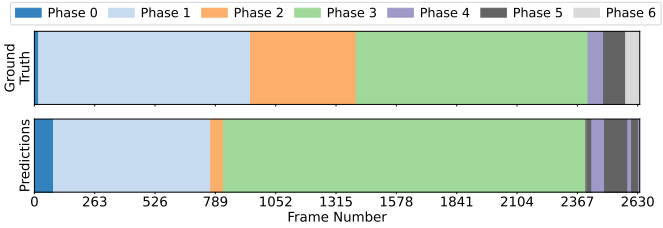


Abbildung A.172: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 76

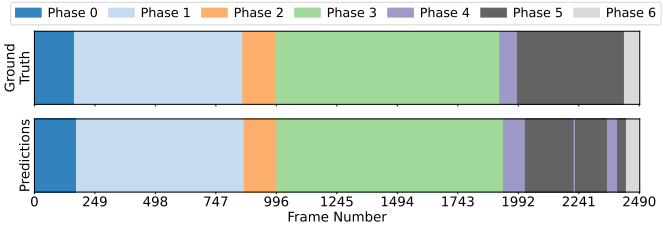


Abbildung A.173: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 77

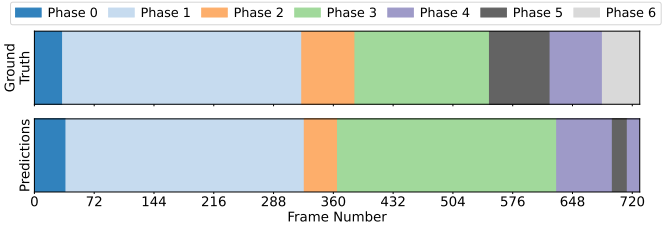


Abbildung A.174: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 78

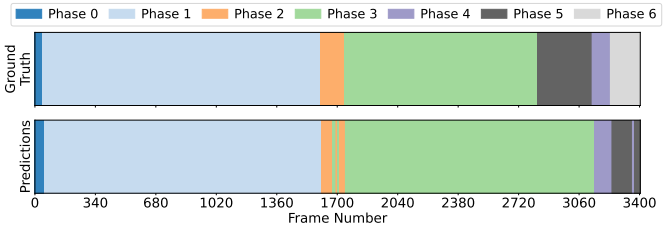


Abbildung A.175: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 79

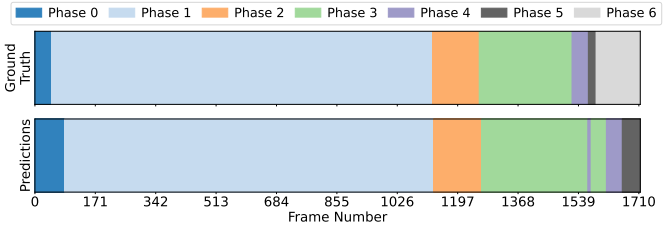


Abbildung A.176: Phasen-Zeit-Diagramm Finetuning auf HeiChole, Cholec80-Video 80

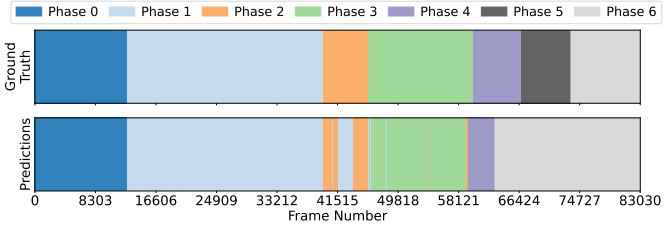


Abbildung A.177: Phasen-Zeit-Diagramm Finetuning auf HeiChole, HeiChole-Video 20

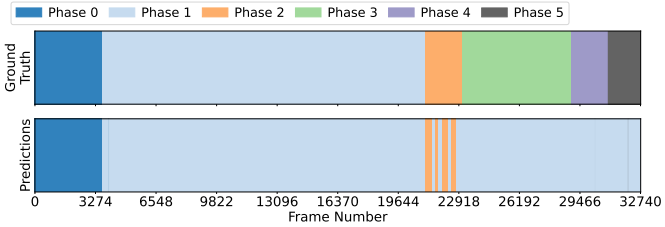


Abbildung A.178: Phasen-Zeit-Diagramm Finetuning auf HeiChole-Video, HeiChole-Video 21

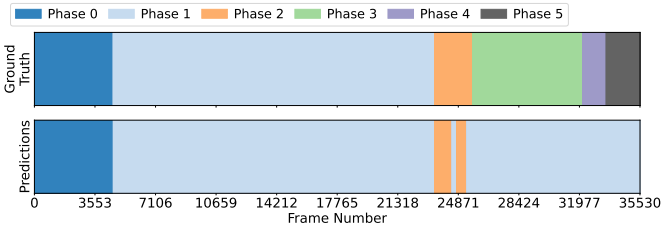


Abbildung A.179: Phasen-Zeit-Diagramm Finetuning auf HeiChole-Video, HeiChole-Video 22

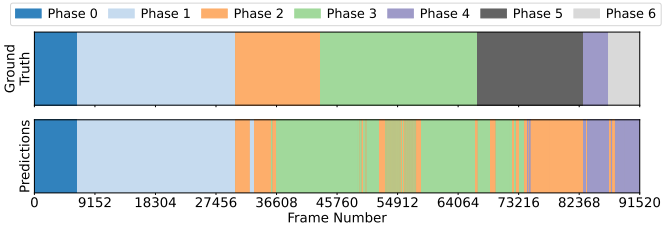


Abbildung A.180: Phasen-Zeit-Diagramm Finetuning auf HeiChole-Video, HeiChole-Video 23

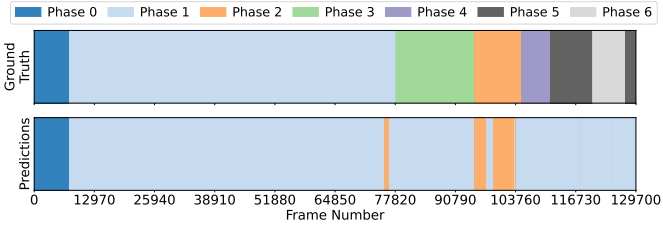


Abbildung A.181: Phasen-Zeit-Diagramm Finetuning auf HeiChole-Video, HeiChole-Video 24

Abbildungsverzeichnis

Tabellenverzeichnis

Eigene Veröffentlichungen

[1] Friedrich Gauger and Lukas Kohout. Projekt TherapyBuilder: Sachbericht zum Verwendungsnachweis, 2022. URL `https://www.tib.eu/de/suchen/id/TIBKAT%3A1881249573`.

[2] Jens Juhl, Korbinian F. Rudolf, Lukas Kohout, Marc B. Schroth, Friedrich Gauger, Christoph Zimmermann, and Wilhelm Stork. Using Augmented Reality and Artificial Intelligence for an Efficient and Safe Preparation of Individual Drug Assortments in Nursing Homes. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6. IEEE, 2022. ISBN 978-1-6654-7095-7. doi: 10.1109/ICECCME55909.2022.9988015.

[3] Lukas Kohout, Manuel Butz, and Wilhelm Stork. Using Acceleration Data for Detecting Temporary Cognitive Overload in Health Care Exemplified Shown in a Pill Sorting Task. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 20–25. IEEE, 2019. ISBN 978-1-7281-2286-1. doi: 10.1109/CBMS.2019.00015.

[4] Lukas Kohout, Jan Scheerer, Christoph Zimmermann, and Wilhelm Stork. Effects of Medical Clothing on Person Re-Identification Algorithms. In *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2022. ISBN 978-1-6654-8299-8. doi: 10.1109/MeMeA54994.2022.9856473.

[5] Markus Lucking, Raphael Manke, Markus Schinle, Lukas Kohout, Stefan Nickel, and Wilhelm Stork. Decentralized patient-centric data management

for sharing IoT data streams. In *2020 International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE, 2020. ISBN 978-1-7281-6371-0. doi: 10.1109/COINS49042.2020.9191653.

[6] Markus Lucking, Esteban Rivera, Lukas Kohout, Christoph Zimmermann, Duygu Polad, and Wilhelm Stork. A video-based vehicle counting system using an embedded device in realistic traffic conditions. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE, 2020. ISBN 978-1-7281-5503-6. doi: 10.1109/WF-IoT48130.2020.9221094.

[7] Markus Scholz, Lukas Kohout, Matthias Horne, Matthias Budde, Michael Beigl, and Moustafa A. Youssef. Device-Free Radio-based Low Overhead Identification of Subject Classes. In Dina Katabi and Archan Misra, editors, *Proceedings of the 2nd workshop on Workshop on Physical Analytics*, pages 1–6, New York, NY, USA, 2015. ACM. ISBN 9781450334983. doi: 10.1145/2753497.2753503.

[8] Marc Schroth, Andreas Ilg, Lukas Kohout, and Wilhelm Stork. A Method for Designing an Embedded Human Activity Recognition System for a Kitchen Use Case Based on Machine Learning. In *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 249–254. IEEE, 2022. ISBN 978-1-6654-5126-0. doi: 10.1109/IAICT55358.2022.9887452.

[9] Christoph Zimmermann and Lukas Kohout. Teilvorhaben: Sensorische Erfassung des Krisenkontextes und Bereitstellung von Interaktionsmechanismen für den Krisenfall : Abschlussbericht zum Verbundvorhaben situCare - Situative Unterstützung und Krisenintervention in der Pflege, 2019. URL `https://www.tib.eu/de/suchen/id/TIBKAT:1681598981/`.

Literaturverzeichnis

[1] Verordnung über das Errichten, Betreiben und Anwenden von Medizinprodukten (Medizinprodukte-Betreiberverordnung): MPBetreibV, 21.8.2002. URL https://www.gesetze-im-internet.de/mpbetreibv/. [zuletzt abgerufen am 08.08.2023].

[2] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a Better Understanding of Context and Context-Awareness. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, and Hans-W. Gellersen, editors, *Handheld and Ubiquitous Computing*, volume 1707 of *Lecture Notes in Computer Science*, pages 304–307. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. ISBN 978-3-540-66550-2. doi: 10.1007/3-540-48157-5_29.

[3] Shay Aharon, Louis-Dupont, Ofri Masad, Kate Yurkova, Lotem Fridman, Lkdci, Eugene Khvedchenya, Ran Rubin, Natan Bagrov, Borys Tymchenko, Tomer Keren, Alexander Zhilko, and Eran-Deci. Super-Gradients, 2021. URL https://zenodo.org/record/7789328.

[4] M. Ahmed and R. Diggory. The correlation between ultrasonography and histology in the search for gallstones. *Annals of the Royal College of Surgeons of England*, 93(1):81–83, 2011. doi: 10.1308/003588411X12851639107070.

[5] AMBOSS GmbH. Laparoskopische Chirurgie, 18.03.2022. URL https://www.amboss.com/de/wissen/Laparoskopische_Chirurgie. [zuletzt abgerufen am 20.12.2022].

[6] AOK-Bundesverband GbR. Wenn Gallensteine Beschwerden machen, 05.01.2023. URL https://www.aok.de/pk/magazin/koerper-

psyche/verdauungssystem/gallensteine-ursache-symptome-und-behandlung/. [zuletzt abgerufen am 12.02.2024].

[7] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, 04.03.2018. URL https://arxiv.org/pdf/1803.01271.pdf. [zuletzt abgerufen am 04.11.2023].

[8] M. Bauer, T. C. Auhuber, R. Kraus, J. Rüggeberg, and Wardmann, K. Müller, P. et al. Glossar perioperativer Prozesszeiten und Kennzahlen. Eine gemeinsame Empfehlung von BDA, BDC, VOPM, VOPMÖ, ÖGARI und SFOPM. *Anästhesiologie & Intensivmedizin*, 61:516–531, 2020. URL https://doi.org/10.19224/ai2020.516. [zuletzt abgerufen am 25.04.2023].

[9] Apurva Bedagkar-Gala and Shishir K. Shah. A Survey of Approaches and Trends in Person Re-Identification. *Image and Vision Computing*, 32(4): 270–286, 2014. ISSN 02628856. doi: 10.1016/j.imavis.2014.02.001.

[10] H.-J. Bender, K. Waschke, and A. Schleppers. Tischlein wechsle dich: Sind Wechselzeiten ein Maß für ein effektives OP-Management? *Anästhesiologie & Intensivmedizin*, 45:529–535, 2004. URL https://www.ai-online.info/images/ai-ausgabe/2004/09-2004/04_09_529-535.pdf. [zuletzt abgerufen am 01.05.2023].

[11] BGBl. Verordnung (EU) 2017/745 über Medizinprodukte: MDR, 2017. URL http://data.europa.eu/eli/reg/2017/745/2023-03-20. [zuletzt abgerufen am 08.08.2023].

[12] Alina Bialkowski, Patrick Lucey, Xinyu Wei, and Sridha Sridharan. Person Re-Identification Using Group Information. In *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2013. ISBN 978-1-4799-2126-3. doi: 10.1109/DICTA.2013.6691512.

[13] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection, 23.04.2020. URL https://arxiv.org/pdf/2004.10934.pdf.

[14] Sebastian Bodenstedt and Martin Wagner. Surgical Workflow and Skill Analysis Challenge (HeiChole Benchmark), 2022.

[15] Bundesministerium für Gesundheit. NAKI - Der Nationale Arbeitskreis zur Implementierung der EU-Verordnungen über Medizinprodukte (MDR) und In-vitro-Diagnostika (IVDR), 2023. URL https://www.bundesgesundheitsministerium.de/themen/gesundheitswesen/medizinprodukte/naki.html. [zuletzt abgerufen am 08.08.2023].

[16] Elisabetta Buscarini, Paolo Tansini, Daniele Vallisa, Alessandro Zambelli, and Luigi Buscarini. EUS for suspected choledocholithiasis: do benefits outweigh costs? A prospective, controlled study. *Gastrointestinal endoscopy*, 57(4):510–518, 2003. ISSN 0016-5107. doi: 10.1067/mge.2003.149.

[17] Thomas Busse. *OP-Management Fibel*. 3M Medica, Neuss, 2 edition, 2009. URL https://www.frankfurt-university.de/fileadmin/standard/Forschung/ZGWR/opmanagementfibel_2auflage_final.pdf. [zuletzt abgerufen am 20.12.2022].

[18] Thomas Busse. OP Barometer 2015, 2016. URL https://www.frankfurt-university.de/fileadmin/standard/Forschung/ZGWR/Praesentation_OP-Barometer_2015_final_quer.pdf. [zuletzt abgerufen am 21.05.2023].

[19] Thomas Busse. OP Barometer 2017, 2018. URL https://www.frankfurt-university.de/fileadmin/standard/Forschung/ZGWR/OP-Barometer_2017.pdf. [zuletzt abgerufen am 21.05.2023].

[20] Thomas Busse. OP Barometer 2019, 2020. URL https://www.frankfurt-university.de/fileadmin/standard/Forschung/ZGWR/OP_Barometer_2019_Aufbereitung_kurz.pdf. [zuletzt abgerufen am 21.05.2023].

[21] Thomas Busse. Zehn Jahre OP-Baromete - eine kritische Bestandsaufnahme. *Management & Krankenhaus*, 39(9):6, 2020. URL https://www.management-krankenhaus.de/management-krankenhaus/management-krankenhaus-ausgabe-9-2020. [zuletzt abgerufen am 07.05.2023].

[22] Canon Inc. Canon VB-H45 Netzwerkkamera: Spezifikationen, 11.07.2018. URL https://www.canon.de/support/consumer/products/network-cameras/vb-series/vb-h45.html?type=manuals. [zuletzt abgerufen am 26.12.2023].

[23] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*, 2017.

[24] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[25] Thomas Carus. *Operationsatlas Laparoskopische Chirurgie: Indikationen - Operationsablauf - Varianten - Komplikationen*. Springer Berlin Heidelberg, Berlin, Heidelberg, 3. aufl. 2014 edition, 2014. ISBN 9783642312465. URL http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1498654.

[26] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level Factorisation Net for Person Re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2109–2118. IEEE, 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00225.

[27] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Qatar Computing Research Institute Alessandro Moschitti, Google Bo Pang, and University of Antwerp Walter Daelemans, editors,

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179.

[28] CMU Perceptual Computing Lab. Repositorium zu OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation, 2016. URL `https://github.com/CMU-Perceptual-Computing-Lab/openpose`. [zuletzt abgerufen am 11.11.2023].

[29] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12263 of *Springer eBook Collection*, pages 343–352. Springer International Publishing and Imprint Springer, Cham, 2020. ISBN 978-3-030-59715-3. doi: 10.1007/978-3-030-59716-0_33.

[30] DailyBigCat. Big Cat Vets - Ginger Serval broke her leg. Getting X-rays. - Part 2 - 7.4.2019, 2019. URL `https://www.youtube.com/watch?v=SMHlKUjTmzg`. [zuletzt abgerufen am 02.12.2023].

[31] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In Cordelia Schmid, editor, *Proceedings / 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, pages 886–893, Los Alamitos, Calif., 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.177.

[32] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In William Cohen and Andrew Moore, editors, *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 233–240, New York, New York, USA, 2006. ACM Press. ISBN 1595933832. doi: 10.1145/1143844.1143874.

[33] Aaron den van Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, 2016.

[34] U. Denzer. Diagnostik bei Cholelithiasis. *Der Gastroenterologe*, 13(1): 23–29, 2018. ISSN 1861-9681. doi: 10.1007/s11377-017-0217-6.

[35] Johannes Diermann. Laparoskopie, 04.2020. URL `https://www.pschyrembel.de/Laparoskopie/K0CJS`. [zuletzt abgerufen am 20.12.2022].

[36] Johannes Diermann. Cholezystektomie, 04.2020. URL `https://www.pschyrembel.de/Cholezystektomie/K04U2`. [zuletzt abgerufen am 20.12.2022].

[37] Johannes Diermann. Laparoskopische Cholezystektomie, 12.2021. URL `https://www.pschyrembel.de/LaparoskopischeCholezystektomie/B056H`. [zuletzt abgerufen am 20.12.2022].

[38] Xinpeng Ding and Xiaomeng Li. Exploring Segment-Level Semantics for Online Phase Recognition From Surgical Videos. *IEEE transactions on medical imaging*, 41(11):3309–3319, 2022. doi: 10.1109/TMI.2022.3182995.

[39] Inga Döbel, Miriam Leis, Manuel Molina Vogelsang, Dmitry Neustroev, Henning Petzka, Stefan Rüping, Angelika Voss, Martin Wegele, and Juliane Welz. Maschinelles Lernen – Kompetenzen, Anwendungen und Forschungsbedarf, 2018. URL `https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/BMBF_Fraunhofer_ML-Ergebnisbericht_Gesamt.pdf`. [zuletzt abgerufen am 31.10.2023].

[40] Sadik Duru. Technik im OP-Saal. In Sadik Duru, Michael Gnant, Klaus Markstaller, and Martin Bodingbauer, editors, *Standards der OP-Patientenlagerung*, pages 111–124. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018. ISBN 978-3-662-57482-9. doi: 10.1007/978-3-662-57483-6_9.

[41] B. Joseph Elmunzer, James M. Scheiman, Glen A. Lehman, Amitabh Chak, Patrick Mosler, Peter D. R. Higgins, Rodney A. Hayward, Joseph Romagnuolo, Grace H. Elta, Stuart Sherman, Akbar K. Waljee, Aparna Repaka, Matthew R. Atkinson, Gregory A. Cote, Richard S. Kwon, Lee McHenry, Cyrus R. Piraka, Erik J. Wamsteker, James L. Watkins, Sheryl J. Korsnes, Suzette E. Schmidt, Sarah M. Turner, Sylvia Nicholson, and Evan L. Fogel. A randomized trial of rectal indomethacin to prevent post-ERCP pancreatitis. *The New England journal of medicine*, 366(15):1414–1422, 2012. doi: 10.1056/NEJMoa1111103.

[42] Wolfgang Ertel. *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. Lehrbuch. Springer Vieweg, Wiesbaden and Heidelberg, 5. auflage edition, 2021. ISBN 978-3-658-32074-4. doi: 10.1007/978-3-658-32075-1.

[43] Chih-Min Fan and Yun-Pei Lu. A Bayesian framework to integrate knowledge-based and data-driven inference tools for reliable yield diagnoses. In Scott J. Mason, editor, *Winter Simulation Conference, 2008*, pages 2323–2329, Piscataway, NJ, 2008. IEEE. ISBN 978-1-4244-2707-9. doi: 10.1109/WSC.2008.4736337.

[44] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. In *2017 IEEE International Conference on Computer Vision*, IEEE Xplore Digital Library, pages 2353–2362, Piscataway, NJ, 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.256.

[45] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[46] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose Dokumentation: Output Format, 22.12.2019. URL `https://github.com/MVIG-SJTU/AlphaPose/blo`

b/master/docs/output.md#output-format. [zuletzt abgerufen am 08.04.2023].

[47] Werner Fleischer. OP-Organisation: Erste Hilfe für das Herzstück. *Dtsch Arztebl International*, 109(50):A–2555–A–2556, 2012. URL https://www.aerzteblatt.de/int/article.asp?id=133666.

[48] Giles M. Foody. Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PloS one*, 18(10):e0291908, 2023. doi: 10.1371/journal.pone.0291908.

[49] Sebastian Freese, Inge Hedemann, Helmut Schmeichel, Martin M. Wilczynski, and Thomas Wille. *Pflegerische Qualitätssicherung im OP: Standardisierte Arbeitsabläufe für den Funktionsdienst*. Kohlhammer Pflegepraxis. Kohlhammer, Stuttgart, 2009. ISBN 9783170190719.

[50] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised Pre-training for Person Re-identification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14745–14754. IEEE, 2021. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.01451.

[51] Yun Fu. *Human Activity Recognition and Prediction*. Springer, Cham and Heidelberg and New York and Dordrecht and London, 2016. ISBN 978-3-319-27002-9. doi: 10.1007/978-3-319-27004-3.

[52] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12904 of *Springer eBook Collection*, pages 593–603. Springer International Publishing and Imprint Springer, Cham, 2021. ISBN 978-3-030-87201-4. doi: 10.1007/978-3-030-87202-1_57.

[53] G. Geldner, L. H. J. Eberhart, S. Trunk, K. G. Dahmen, T. Reissmann, T. Weiler, and A. Bach. Effizientes OP-Management Vorschläge zur Optimierung von Prozessabläufen als Grundlage für die Erstellung eines OP-Statuts. *Der Anaesthesist*, 51(9):760–767, 2002. ISSN 0003-2417. doi: 10.1007/s00101-002-0362-1.

[54] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: continual prediction with LSTM. *Neural computation*, 12(10):2451–2471, 2000. ISSN 0899-7667. doi: 10.1162/089976600300015015.

[55] Roger Gfrörer. *Das Operationsteam: Eine Analyse der Verhältnisse der Zusammenarbeit im Operationssaal*. Gesundheits- und Qualitätsmanagement. Gabler Verlag, s.l., 1. aufl. edition, 2008. ISBN 9783835009219. URL `http://gbv.eblib.com/patron/FullRecord.aspx?p=748724`.

[56] Vanja Giljaca, Kurinchi Selvan Gurusamy, Yemisi Takwoingi, David Higgie, Goran Poropat, Davor Štimac, and Brian R. Davidson. Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones. *The Cochrane database of systematic reviews*, 2015 (2):CD011549, 2015. doi: 10.1002/14651858.CD011549.

[57] Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448. IEEE, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.169.

[58] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR 2014*, pages 580–587, Los Alamitos, California, 2014. IEEE Computer Society. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.81.

[59] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. Advances in in computer vision and pattern recognition. Springer, London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN 978-0262035613. URL http://www.deeplearningbook.org. [zuletzt abgerufen am 31.10.2023].

[61] Douglas Gray, Shane Brennan, and Hai Tao, editors. *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking*, 2007.

[62] Carsten N. Gutt and Holger Listle. Laparoskopische Cholezystektomie. In Tobias Keck and Christoph T. Germer, editors, *Minimalinvasive Viszeralchirurgie*, pages 123–136. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-53203-4. doi: 10.1007/978-3-662-53204-1_13.

[63] Heinz Haferkorn. *Optik: Physikalisch-technische Grundlagen und Anwendungen*. Wiley-VCH, Weinheim, 4., bearb. und erw. aufl. edition, 2003. ISBN 9783527403721. doi: 10.1002/3527609032.

[64] Barbara Hammer and Jochen J. Steil. Tutorial: Perspectives on Learning with RNNs. In M. Verleysen, editor, *Proc. European Symposium Artificial Neural Networks*, pages 357–368. D-side publication, 2002.

[65] Thomas Harzenetter. *Cholezystektomie am Kreiskrankenhaus Wasserburg/Inn: Indikation und Behandlungsergebnisse bei 241 Patienten*. Dissertation, Technische Universität München, München, 2005. URL https://mediatum.ub.tum.de/doc/602639/602639.pdf. [zuletzt abgerufen am 21.12.2022].

[66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN, 20.03.2017. URL https://arxiv.org/pdf/1703.06870.pdf.

[67] Ekbert Hering and Rolf Martin, editors. *Optik für Ingenieure und Naturwissenschaftler: Grundlagen und Anwendungen : mit zahlreichen Bildern, Tabellen, Beispielen*. Fachbuchverlag Leipzig im Carl Hanser Verlag, München, 2017. ISBN 3446442812. doi: 44281. URL http://www.hanser-fachbuch.de/9783446442818.

[68] HIMSS EUROPE. Auf den Spuren der Zeitdiebe im Krankenhaus: Die wahre Belastung durch Dokumentation an deutschen Akutkrankenhäusern wird

unterschätzt, 19.03.2015. URL https://www.dragon-speaking.de/download/HIMSS-Europe-Studie.pdf?m=1434964003&. [zuletzt abgerufen am 19.05.2023].

[69] Sepp Hochreiter. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998. ISSN 0218-4885. doi: 10.1142/S0218488598000094.

[70] Sepp Hochreiter and Jürgen Schmidhuber. LSTM can Solve Hard Long Time Lag Problems. In M. C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/a4d2f0d23dcc84ce983ff9157f8b7f88-Paper.pdf. [zuletzt abgerufen am 04.11.2023].

[71] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[72] Ronghang Hu and Amanpreet Singh. UniT: Multimodal Multitask Learning with a Unified Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1419–1429. IEEE, 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00147.

[73] Quan Hua. [CVPR 2019] Efficient Online Multi-Person 2D Pose Tracking with Recurrent Spatio-Temporal Affinity Fields: Real-time detect and track 2D poses of multiple people at 30 fps on a single GPU., 2019. URL https://towardsdatascience.com/cvpr-2019-efficient-online-multi-person-2d-pose-tracking-with-recurrent-spatio-temporal-affinity-25c4914e5f6. [zuletzt abgerufen am 09.11.2023].

[74] S. Hunziker, A. Baumgart, C. Denz, and G. Schüpfer. Ökonomischer Nutzen der überlappenden Einleitung: Untersuchung mithilfe eines Computersimulationsmodells. *Der Anaesthesist*, 58(6):623–632, 2009. ISSN

0003-2417. doi: 10.1007/s00101-009-1551-y. URL https://link.springer.com/article/10.1007/s00101-009-1551-y. [zuletzt abgerufen am 20.12.2022].

[75] Katsushi Ikeuchi. *Computer Vision: A Reference Guide*. Springer reference. Springer, New York and Heidelberg, 1 edition, 2014. ISBN 978-0-387-30771-8. doi: 10.1007/978-0-387-31439-6. URL https://link.springer.com/referencework/10.1007/978-0-387-31439-6. [zuletzt abgerufen am 31.10.2023].

[76] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schieke. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision (ECCV)*, 2016.

[77] Institut für das Entgeltsystem im Krankenhaus GmbH. InEK Datenbrowser: Datenlieferung DRG 2021 gruppiert nach 2022, 2022. URL https://datenbrowser.inek.org/. [zuletzt abgerufen am 29.12.2022].

[78] International Society for Computer Aided Surgery (ISCAS). Endoscopic Vision Challenge: A MICCAI Challenge, 2023. URL http://opencas.dkfz.de/endovis/. [zuletzt abgerufen am 18.05.2023].

[79] Karl Jähn and Wolfgang Cibis. Krankenhausinformationssystem, 04.2016. URL https://www.pschyrembel.de/Krankenhausinformationssystem/K0QP7. [zuletzt abgerufen am 11.02.2023].

[80] Matthias Janda, Andreas Brosin, and Daniel A. Reuter. Modernes OP-Management an einem Haus der Maximalversorgung. *Unfallchirurgie (Heidelberg, Germany)*, 125(10):811–820, 2022. doi: 10.1007/s00113-022-01222-8.

[81] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. *IEEE transactions on medical imaging*, 37(5):1114–1126, 2018. doi: 10.1109/TMI.2017.2787657.

[82] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical image analysis*, 59: 101572, 2020. doi: 10.1016/j.media.2019.101572.

[83] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal Memory Relation Network for Workflow Recognition From Surgical Video. *IEEE transactions on medical imaging*, 40(7):1911–1923, 2021. doi: 10.1109/TMI.2021.3069471.

[84] Glenn Jocher and Ayush Chaurasia. Ultralytics Dokumentation, 2023. URL `https://docs.ultralytics.com/`. [zuletzt abgerufen am 20.11.2023].

[85] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL `https://github.com/ultralytics/ultralytics`. [zuletzt abgerufen am 02.10.2023].

[86] Michael I. Jordan. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 531–546. Hillsdale, NJ: Erlbaum, 1986.

[87] KARL STORZ SE & Co. KG. OR1, 2023. URL `https://www.karlstorz.com/de/de/karl-storz-or1.htm`. [zuletzt abgerufen am 21.05.2023].

[88] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR 2014*, pages 1725–1732, Los Alamitos, California, 2014. IEEE Computer Society. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.223.

[89] H. G. Kenngott, M. Wagner, A. A. Preukschas, and B. P. Müller-Stich. Der intelligente Operationssaal : Vom passiven Gerätepark zum mitdenkenden, kognitiven Assistenten. *Der Chirurg; Zeitschrift fur alle Gebiete der operativen Medizen*, 87(12):1033–1038, 2016. doi: 10.1007/s00104-016-0308-

9. URL https://link.springer.com/article/10.1007/s00104-016-0308-9. [zuletzt abgerufen am 21.05.2023].

[90] Michael Knauthe. *Effizienzsteigerung im Operationsbereich durch ein Prozessmanagement*. Dissertation, Ernst-Moritz-Arndt-Universität, Greifswald, 2006. URL https://nbn-resolving.org/urn:nbn:de:gbv:9-200510-8. [zuletzt abgerufen am 22.04.2023].

[91] Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO). Anforderungen an die Hygiene bei der Aufbereitung von Medizinprodukten. Empfehlung der Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO) beim Robert Koch-Institut (RKI) und des Bundesinstitutes für Arzneimittel und Medizinprodukte (BfArM). *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 55(10):1244–1310, 2012. doi: 10.1007/s00103-012-1548-6.

[92] Yu Kong and Yun Fu. Human Action Recognition and Prediction: A Survey, 29.06.2018. URL https://arxiv.org/pdf/1806.11230.pdf.

[93] Peter Kornprat, Georg Werkgartner, Herwig Cerwenka, Heinz Bacher, Azab El-Shabrawi, Peter Rehak, and Hans Jörg Mischinger. Prospective study comparing standard and robotically assisted laparoscopic cholecystectomy. *Langenbeck's archives of surgery*, 391(3):216–221, 2006. ISSN 1435-2443. doi: 10.1007/s00423-006-0046-4.

[94] Michael Kranzfelder, Armin Schneider, Adam Fiolka, Sebastian Koller, Silvano Reiser, Thomas Vogel, Dirk Wilhelm, and Hubertus Feussner. Reliability of sensor-based real-time workflow recognition in laparoscopic cholecystectomy. *International journal of computer assisted radiology and surgery*, 9(6):941–948, 2014. doi: 10.1007/s11548-014-0986-z.

[95] Kirstin Kraus. *Klinische Pfade: Logistische Integration der Termin- und Kapazitätsplanung*. Diplomarbeit, Universität Saarbrücken, Saarbrücken, 2007. URL https://silo.tips/download/abbildungsverzeichnisiii-tabellenverzeichnisiv-abkrzungsverzeichnis-v-1-einfhrun. [zuletzt abgerufen am 21.12.2022].

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc, 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[97] Ben Kröse, Tim van Oosterhout, and Tim van Kasteren. Activity Monitoring Systems in Health Care. In Albert Ali Salah and Theo Gevers, editors, *Computer Analysis of Human Behavior*, pages 325–346. Springer London, London, 2011. ISBN 978-0-85729-993-2. doi: 10.1007/978-0-85729-994-9_12.

[98] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6848-6. doi: 10.1007/978-1-4614-6849-3.

[99] Michael Langhorst, Enno Bialas, Julia Katharina Bergmann, and Jörg Ulrich Ansorg. Prozessdaten im OP am Beispiel der Cholecystektomie. *Passion Chirurgie*, 5(1), 2015. URL https://www.bdc.de/prozessdaten-im-op-am-beispiel-der-cholecystektomie-3. [zuletzt abgerufen am 21.12.2022].

[100] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal Convolutional Networks for Action Segmentation and Detection. In *30th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1003–1012, Piscataway, NJ, 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.113.

[101] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[102] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. YOLOv6: A Single-Stage Object Detection Framework

for Industrial Applications, 07.09.2022. URL https://arxiv.org/pdf/2209.02976.pdf.

[103] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[104] Wei Li, Rui Zhao, and Xiaogang Wang. Human Reidentification with Transferred Metric Learning. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, volume 7724 of *Lecture Notes in Computer Science*, pages 31–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-37330-5. doi: 10.1007/978-3-642-37331-2_3.

[105] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *CVPR 2014*, pages 152–159, Los Alamitos, California, 2014. IEEE Computer Society. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.27.

[106] Margret Liehn and Hannelore Schlautmann. *1×1 der chirurgischen Instrumente*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-53956-9. doi: 10.1007/978-3-662-53957-6.

[107] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer vision - ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer, Cham, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0_2.

[108] David G. Lowe. Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE International Conference on Computer Vision*, pages 1150–1157, Los Alamitos, Calif., 1999. IEEE COMPUTER SOC. ISBN 0-7695-0164-8. doi: 10.1109/ICCV.1999.790410.

[109] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206827.

[110] Wenyu Lv, Yian Zhao, Shangliang Xu, Jinman Wei, Guanzhong Wang, Cheng Cui, Du Yuning, Qingqing Dang, and Yi Liu. DETRs Beat YOLOs on Real-time Object Detection, 17.04.2023. URL `https://arxiv.org/pdf/2304.08069.pdf`.

[111] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-Nötzel, Hannes G. Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Tobias Roß, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. Müller-Stich, Pierre Jannin, and Stefanie Speidel. Surgical data science - from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022. doi: 10.1016/j.media.2021.102306.

[112] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge Univ. Press, Cambridge, reprinted. edition, 2009. ISBN 9780521865715.

[113] Medizinprodukte Journal. Der NAKI: Struktur, Zusammensetzung und Aufgaben. *Medizinprodukte Journal*, 25(1):25–29, 2018.

URL https://www.medizinprodukte-journal.de/wp-content/uploads/2018/02/MPJ_1_2018_Beitrag_NAKI_web.pdf. [zuletzt abgerufen am 08.08.2023].

[114] Lawrence Mosley. *A balanced approach to the multi-class imbalance problem*. Dissertation, Iowa State University, Ames, Iowa, 2013. URL https://dr.lib.iastate.edu/handle/20.500.12876/27724. [zuletzt abgerufen am 14.02.2024].

[115] Hirenkumar Nakawala, Roberto Bianchi, Laura Erica Pescatori, Ottavio de Cobelli, Giancarlo Ferrigno, and Elena de Momi. "Deep-Onto" network for surgical workflow and context recognition. *International journal of computer assisted radiology and surgery*, 14(4):685–696, 2019. doi: 10.1007/s11548-018-1882-8.

[116] Andrew Ng. Machine Learning Yearning: Technical Strategy for AI Engineers, in the Era of Deep Learning, 2016. URL https://github.com/ajaymache/machine-learning-yearning/tree/master. [zuletzt abgerufen am 31.10.2023].

[117] Guanghan Ning, Jian Pei, and Heng Huang. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4456–4465. IEEE, 2020. ISBN 978-1-7281-9360-1. doi: 10.1109/CVPRW50498.2020.00525.

[118] Christopher Olah. Understanding LSTM Networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [zuletzt abgerufen am 09.12.2023].

[119] Falk Osterloh. Krankenhäuser: DRG-System auf dem Prüfstand. *Dtsch Arztebl International*, 119(40):A3252–3258, 2022. URL https://www.aerzteblatt.de/archiv/227891/Krankenhaeuser-DRG-System-auf-dem-Pruefstand.

[120] Nicolas Padoy. *Workflow and Activity Modeling for Monitoring Surgical Procedures*. Dissertation, Université Henri Poincaré - Nancy 1 and Technische Universität München, 2010. URL https://theses.hal.science/tel-01748567. [zuletzt abgerufen am 27.12.2022].

[121] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[122] David M. W. Powers. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *International Journal of Machine Learning Technology*, 2, 2011. URL https://arxiv.org/pdf/2010.16061.pdf.

[123] Pschyrembel Redaktion. perioperativ, 03.2019. URL https://www.pschyrembel.de/perioperativ/B0MFG. [zuletzt abgerufen am 20.12.2022].

[124] Bharath Raj and Yoni Osin. An Overview of Human Pose Estimation with Deep Learning, 2019. URL https://medium.com/beyondminds/an-overview-of-human-pose-estimation-with-deep-learning-d49eb656739b. [zuletzt abgerufen am 09.11.2023].

[125] Jörg Ratzsch. OP-Material im Körper vergessen: Laut eines Gutachten gab es im vergangenen Jahr 84 Todesfälle und rund 2.700 Gesundheitsschäden durch Behandlungsfehler. *Badische Neueste Nachrichten*, 2023(190):3, 18.08.2023.

[126] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, 09.04.2018. URL https://arxiv.org/pdf/1804.02767.pdf.

[127] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger, 25.12.2016. URL https://arxiv.org/pdf/1612.08242.pdf.

[128] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection, 08.06.2015. URL https://arxiv.org/pdf/1506.02640.pdf.

[129] Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multi-modal uniform deep learning for RGB-D person re-identification. *Pattern Recognition*, 72:446–457, 2017. ISSN 00313203. doi: 10.1016/j.patcog.2017.06.037.

[130] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press.

[131] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[132] Richard Wolf GmbH. core nova: Fokussiert. Unkompliziert. Integriert., 2023. URL https://www.richard-wolf.com/de/loesungen/integration. [zuletzt abgerufen am 21.05.2023].

[133] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In Gang Hua and Hervé Jégou, editors, *Computer vision - ECCV 2016 Workshops*, volume 9914 of *Lecture Notes in Computer Science*, pages 17–35. Springer, Cham, 2016. ISBN 978-3-319-48880-6. doi: 10.1007/978-3-319-48881-3_2.

[134] RIWOlink GmbH. core.nova: Versatile. User-friendly. Integrated., 2023. URL https://www.riwolink.com/en/solutions/integration. [zuletzt abgerufen am 21.05.2023].

[135] Jelle P. Ruurda, Paul L. Visser, and Ivo A. M. J. Broeders. Analysis of procedure time in robot-assisted surgery: comparative study in laparoscopic

cholecystectomy. *Computer aided surgery*, 8(1):24–29, 2003. doi: 10.3109/10929080309146099.

[136] Riccardo Satta. Appearance Descriptors for Person Re-identification: a Comprehensive Review, 22.07.2013. URL `https://arxiv.org/pdf/1307.5748.pdf`.

[137] B. Schilit, N. Adams, and R. Want. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*, pages 85–90. IEEE, 1994. ISBN 978-0-7695-3451-0. doi: 10.1109/WMCSA.1994.16.

[138] Anna-Elisa Schulze-Schleithoff and Pschyrembel Redaktion. Gallenblase, 10.2018. URL `https://www.pschyrembel.de/Gallenblase/K08CR`. [zuletzt abgerufen am 21.12.2022].

[139] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*, 2017.

[140] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos, 09.06.2014. URL `https://arxiv.org/pdf/1406.2199.pdf`.

[141] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[142] Sandro Skansi. *Introduction to deep learning: From logical calculus to artificial intelligence*. Undergraduate topics in computer science. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-73003-5. doi: 10.1007/978-3-319-73004-2.

[143] Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation, 24.08.2018. URL https://arxiv.org/pdf/1808.08180.pdf. [zuletzt abgerufen am 02.12.2023].

[144] Statistisches Bundesamt (Destatis). Fallpauschalenbezogene Krankenhausstatistik (DRG-Statistik): Diagnosen, Prozeduren, Fallpauschalen und Case Mix der vollstationären Patientinnen und Patienten in Krankenhäusern, 2016. URL https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Publikationen/Downloads-Krankenhaeuser/fallpauschalen-krankenhaus-2120640167004.pdf?__blob=publicationFile. [zuletzt abgerufen am 29.12.2022].

[145] Statistisches Bundesamt (Destatis). Krankenhäuser: Einrichtungen, Betten und Patientenbewegung, 2022. URL https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Tabellen/gd-krankenhaeuser-jahre.html. [zuletzt abgerufen am 21.05.2023].

[146] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11208 of *Lecture Notes in Computer Science*, pages 501–518. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01224-3. doi: 10.1007/978-3-030-01225-0_30.

[147] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1–9, Piscataway, NJ, 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298594.

[148] Juan Terven and Diana Cordova-Esparza. A Comprehensive Review of YOLO: From YOLOv1 and Beyond, 02.04.2023. URL https://arxiv.org/pdf/2304.00501.pdf.

[149] The MathWorks, Inc. Introducing Deep Learning with MATLAB, 2021. URL https://de.mathworks.com/campaigns/offers/deep-learning-with-matlab.html. [zuletzt abgerufen am 12.02.2024].

[150] O. Tschudi, G. Schüpfer, M. Bauer, and R. M. Waeschle. Effiziente Nutzung von OP-Kapazitäten–das Luzerner Konzept. Eine Methodenbeschreibung. *Anästhesiologie & Intensivmedizin*, 58:85–93, 2017. URL https://www.ai-online.info/archiv/2017/02-2017/effiziente-nutzung-von-op-kapazitaeten-das-luzerner-konzept-eine-methodenbeschreibung.html. [zuletzt abgerufen am 24.04.2023].

[151] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957.

[152] Ultralytics Inc. Ultralytics YOLOv5, 2020. URL https://github.com/ultralytics/yolov5. [zuletzt abgerufen am 20.11.2023].

[153] Ultralytics Inc. Ultralytics Homepage, 2023. URL https://www.ultralytics.com/. [zuletzt abgerufen am 20.11.2023].

[154] C. J. van Rijsbergen. *Information retrieval.* Butterworth, London, 2. ed., repr edition, 1981. ISBN 0408709294.

[155] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 12.06.2017. URL https://arxiv.org/pdf/1706.03762.pdf.

[156] Verband der Ersatzkassen e. V. (vdek). Krankenhausfinanzierung, 2021. URL https://www.vdek.com/vertragspartner/Krankenh

aeuser/krankenhausfinanzierung.html. [zuletzt abgerufen am 13.05.2023].

[157] Claudia Vollmert. *Die Single-Incision-Cholezystektomie:eine "narbenlose" Alternative zur laparoskopischen Cholezystektomie*. Dissertation, Eberhard Karls Universität Tübingen, Tübingen, 2015. URL http://dx.doi.org/10.15496/publikation-7417. [zuletzt abgerufen am 29.12.2022].

[158] R. M. Waeschle, J. Hinz, F. Bleeker, B. Sliwa, A. Popov, C. E. Schmidt, and M. Bauer. Mythos OP-Minute : Leitfaden zur Kalkulation von DRG-Erlösen pro Op-Minute. *Der Anaesthesist*, 65(2):137–147, 2016. ISSN 0003-2417. doi: 10.1007/s00101-015-0124-5.

[159] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Carissa Reid, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Long Zhao, Zhenxiao Ge, Haiming Sun, Di Xie, Mengqi Guo, Daochang Liu, Hannes G. Kenngott, Felix Nickel, Moritz von Frankenberg, Franziska Mathis-Ullrich, Annette Kopp-Schneider, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. *Medical image analysis*, 86:102770, 2023. doi: 10.1016/j.media.2023.102770.

[160] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 06.07.2022. URL https://arxiv.org/pdf/2207.02696.pdf.

[161] Kejun Wang, Haolin Wang, Meichen Liu, Xianglei Xing, and Tian Han. Survey on person re–identification based on deep learning. *CAAI Transactions on Intelligence Technology*, 3(4):219–227, 2018. ISSN 2468-2322. doi: 10.1049/trit.2018.1001.

[162] Jin-Mao Wei, Xiao-Jie Yuan, Qing-Hua Hu, and Shu-Qin Wang. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37(5): 3799–3809, 2010. ISSN 09574174. doi: 10.1016/j.eswa.2009.11.040.

[163] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88. IEEE, 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00016.

[164] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[165] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dąbrowski. On the Unreasonable Effectiveness of Centroids in Image Retrieval. In Teddy Mantoro, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto, editors, *Neural Information Processing*, volume 13111 of *Lecture Notes in Computer Science*, pages 212–223. Springer International Publishing, Cham, 2021. ISBN 978-3-030-92272-6. doi: 10.1007/978-3-030-92273-3_18.

[166] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

[167] Fangqiu Yi, Yanfeng Yang, and Tingting Jiang. Not End-to-End: Explore Multi-Stage Architecture for Online Surgical Phase Recognition. In Lei Wang, Juergen Gall, Tat-Jun Chin, Imari Sato, and Rama Chellappa, editors, *Computer vision - ACCV 2022*, volume 13844 of *Lecture Notes in Computer*

Science, pages 417–432. Springer, Cham, 2023. ISBN 978-3-031-26315-6. doi: 10.1007/978-3-031-26316-3_25.

[168] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-Grained Person Re-identification. *International Journal of Computer Vision*, 128(6):1654–1672, 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01259-0.

[169] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching, 2017.

[170] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future, 10.10.2016. URL `https://arxiv.org/pdf/1610.02984.pdf`.

[171] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.133.

[172] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *2017 IEEE International Conference on Computer Vision*, IEEE Xplore Digital Library, pages 3774–3782, Piscataway, NJ, 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.405.

[173] Kaiyang Zhou and Tao Xiang. Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch, 2019.

[174] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3701–3711. IEEE, 2019. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00380.

[175] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning Generalisable Omni-Scale Representations for Person Re-Identification.

IEEE Transactions on Pattern Analysis and Machine Intelligence, PP, 2021. doi: 10.1109/TPAMI.2021.3069237.