Privacy-Friendliness in Human Behavior Analysis for Urban Surveillance Scenarios

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie KIT

genehmigte

Dissertation

von

Thomas Christoph Golda, M.Sc.

aus Schweinfurt

Tag der mündlichen Prüfung:	03.12.2024
Erster Gutachter:	Prof. Dr.-Ing. habil. Jürgen Beyerer
Zweiter Gutachter:	Prof. Emanuel Aldea

Abstract

This thesis investigates the integration of privacy-friendly human behavior analysis within urban video surveillance systems, addressing the prevalent privacy concerns associated with human-related video and image data. The central approach involves the use of human pose information, representing pedestrians as skeletons, which effectively encapsulates necessary data for behavior analysis while maintaining privacy. Despite numerous algorithms and machine learning models for human pose estimation, their application in urban video surveillance remains challenging and underexplored. To address this, a synthetic dataset tailored to typical urban surveillance scenarios was developed, featuring multiple pedestrians and mutual occlusions. A new set of metrics, the Graph Crowd Index, was introduced to assess crowdedness, improving upon the existing Crowd Index for video surveillance contexts. To bridge the domain gap between synthetic and real-world data, a Cycle-GAN-based model was employed for domain adaptation. Experiments indicated that a well-designed target domain representation enhances human pose estimation performance, though the selection of target domain data is crucial.

Following these first investigated aspects, the thesis takes a look on behavior analysis using holistic and human-centered methods, and compares them against state-of-the-art approaches. A unified evaluation pipeline was introduced to benchmark various skeleton-based behavior analysis methods, highlighting the difficulty of maintaining accuracy while ensuring privacy. This pipeline facilitated the comparison of state-of-the-art methods, underlining the challenges and potential of privacy-preserving models. The human-centered DualHeadAE model presented in this thesis demonstrated competitive performance, outperforming the developed holistic approach MurzGAN. The latter proved to work in simpler academic scenarios but

struggled with more complex and dynamic environments. The DualHeadAE approach showed promising results but indicated areas for further improvement, such as incorporating further domain knowledge to enhance performance in challenging cases even further.

In conclusion, this thesis contributes to privacy-aware methodologies in smart video surveillance, presenting a behavior analysis pipeline suitable for real-world applications. While there is room for improvement, the developed methods can serve as semi-automatic pre-filtering steps for authorities, advancing towards more focused classification models. The findings demonstrate the feasibility of preserving data privacy in modern computer vision systems, though their application remains governed by regulations such as the General Data Protection Regulation and local laws.

Kurzfassung

Diese Dissertation untersucht die Integration von Datenschutzaspekten in den Prozess der Verhaltensanalyse von Menschen für den Einsatz von Videoüberwachungssysteme im städtischen Raum und befasst sich mit den weit verbreiteten Datenschutzbedenken im Zusammenhang mit menschbezogenen Video- und Bilddaten. Der zentrale Ansatz der in der Arbeit verfolgt wird umfasst die Nutzung von menschlichen Poseninformationen, wobei Fußgänger als Skelette dargestellt werden, die die notwendigen Daten zur Verhaltensanalyse effektiv erfassen und gleichzeitig die Privatsphäre wahren. Trotz der Existenz zahlreicher Ansätze zur Schätzung menschlicher Posen bleibt deren Anwendung in der städtischen Videoüberwachung herausfordernd und wenig erforscht. Um dieses Thema zu adressieren, wurde ein synthetischer Datensatz entwickelt, der auf typische städtische Überwachungsszenarien zugeschnitten ist und gleichzeitig viele Personen zeigt, mit Szenen die insbesondere Herausforderungen umfassen wie die gegenseitige Verdeckungen von Personen. Ein neuer Satz von Metriken, der Graph Crowd Index, wurde eingeführt, um die Dichte von Menschenmengen zu bewerten und den bestehenden Crowd Index für Videoüberwachungskontexte zu verbessern. Um die Lücke zwischen synthetischen und realen Daten zu überbrücken, wurde ein Cycle-GAN-basiertes Modell für die Anpassung der Datendomänen entwickelt. Experimente zeigten, dass gut gestaltetene Zieldomänendaten die Leistung der menschlichen Posenschätzer sichtlich verbessert, wobei die geeignete und passende Auswahl der Zieldomänendaten entscheidend ist.

Nach diesen ersten untersuchten Aspekten setzt sich die Dissertation im zweiten inhaltlichen Teil mit der Verhaltensanalyse unter Verwendung holistischer und menschenzentrierter Methoden auseinander und vergleicht sie mit aktuellen Ansätzen die den Stand der Technik bilden. Um diese Verfahren

miteinander vergleichen zu können, wird eine vereinheitlichte Evaluationspipeline eingeführt, um verschiedene posenbasierte Verhaltensanalyse-Ansätze zu benchmarken. Diese Pipeline erleichtert den Vergleich von aktuellen Methoden und unterstreicht die Herausforderungen sowie das Potenzial datenschutzfreundlicher Modelle. Das menschenzentrierte DualHeadAE-Modell, das in dieser Arbeit vorgestellt wird, zeigte eine wettbewerbsfähige Leistung und übertrifft den entwickelten holistischen Ansatz MurzGAN. Der DualHeadAE-Ansatz zeigte vielversprechende Ergebnisse, deutete jedoch auf Bereiche hin, in denen eine weitere Verbesserung möglich ist, beispielsweise durch die Einbeziehung weiteres Domänenwissens zur weiteren Steigerung der Erkennungsleistung in besonders herausfordernden Fällen.

Abschließend trägt diese Dissertation zu datenschutzfreundlichen Methoden in der intelligenten Videoüberwachung bei und präsentiert eine Verhaltensanalyse-Pipeline, die für die Bewertung realer Anwendungen geeignet ist. Obwohl es noch Raum für Verbesserungen gibt, können die entwickelten Methoden bereits als halbautomatische Vorfilterungsschritte für Behörden dienen und den Weg zu fokussierteren Klassifikationsmodellen ebnen. Die Ergebnisse zeigen auf, dass die Berücksichtigung von menschlicher Privatsphäre in modernen Videoauswertesystemen umsetzbar ist, solche Systeme jedoch weiterhin vor großen Herausforderungen stehen, insbesondere da ihre Anwendung durch Gesetze wie den Artificial Intelligence Act der Europäischen Union oder die Datenschutzgrundverordnung bzw. ihr europäisches Pendant die General Data Protection Regulation, sowie weitere lokale Gesetze strikt geregelt oder eingeschränkt bleibt.

Acknowledgements

In the winding journey of research, there are guiding lights and supportive hands that illuminate the path and lend strength to the weary traveler. To those who have played integral roles in shaping my academic pursuit, I extend my deepest gratitude.

First and foremost, I am indebted to *Prof. Jürgen Beyerer*, whose wisdom and guidance have been the cornerstone of my research endeavors. The annual Sommerseminar in Triberg, nestled in the serene embrace of the Black Forest, provided not just a forum for academic discourse but a nurturing environment where ideas flourished and insights were shared. His unwavering commitment to the progress of his students has been a source of inspiration. I extend heartfelt appreciation to *Prof. Emanuel Aldea*, whose expertise and discerning eye have enriched my work as a second reviewer. Our paths converged during a cross-cultural research odyssey, where the fusion of French and German intellects gave birth to innovative approaches in crowd monitoring. Though our interactions were tethered to academia, his willingness to lend his expertise speaks volumes of his dedication to fostering scholarly discourse.

To my students, both past and present, I offer collective gratitude for their invaluable contributions. Whether through their theses or as collaborators in the trenches of implementation, each encounter has left an indelible mark on my academic journey. Your enthusiasm and dedication have been the catalysts for growth and innovation.

Last but certainly not least, I extend boundless appreciation to my girlfriend and family. Their unwavering support, boundless patience, and unwavering

belief in my pursuits have been the bedrock upon which I've built my motivation. In moments of doubt and struggle, their love and encouragement have been a guiding light, illuminating even the darkest of paths.

To all who have played a part, whether great or small, in this academic odyssey, I offer my sincerest thanks. Your contributions have not only enriched my work but have also enriched my life in ways beyond measure.

Contents

Appendix

Notation

This chapter introduces the notation and symbols which are used in this thesis.

General notation

Scalars	italic Roman and Greek lowercase letters	x, α
Sets	italic Roman uppercase letters	E
Vectors	bold Roman lowercase letters	$\mathbf{t}$
Vector entry	italic Roman lowercase letters with raised index in parentheses	$t^{(i)}$
Matrices	bold Roman uppercase letters	$\mathbf{R}$
Matrix entry	italic Roman lowercase letters with raised indices in parentheses	$r^{(i,j)}$
State spaces	bold calligraphic Roman uppercase letters	$\boldsymbol{\mathcal{X}}$
Multi-dimensional random variables	bold italic Roman uppercase letters	$\boldsymbol{E}$
Distributions	calligraphic uppercase letters	$\mathcal{N}$

In multidimensional sets of elements related to time series, the first index denotes time.

Distributions

$\mathcal{N}$	Gaussian normal distribution
χ^2_n	n-dimensional chi-square distribution

Numbers and indexing

$\mathbb{N}$	natural numbers
$\mathbb{N}_0$	natural numbers including zero (non-negative integers)
k, t	discrete points in time
i, j, ℓ, q	indexing for objects, measurements and points
$\mathbb{1}$	matrix of ones / all-ones matrix
$\mathbf{0}$	vector of zeros
$\mathbf{I}$	identity matrix

Neural Networks

Θ	parameters of a given neural network, including weights $\mathbf{W}$ and biases $\mathbf{b}$
ϕ	activation function
$\mathcal{L}$	loss function

1 Introduction

1.1 Motivation

In today's world, the ubiquity of electronic devices has become a near-universal experience for almost everyone. Individuals from various generations, including those born in the 80s and 90s, have been significantly influenced by the numerous technological advancements of our modern era. The widespread success and accessibility of the internet, as well as electronic devices such as music players, gaming consoles, personal computers, and digital photography, serve as just a few examples of how technological progress has shaped humanity.

The Digital Revolution has had a profound impact on society, to the extent that many children today have grown up never knowing a life without constant internet connectivity and digital technology. Particularly in the 21st century, remarkable advancements in these fields have resulted in even the simplest devices and concepts incorporating complex electrical circuits and algorithms.

This transforming development has given rise to a globalized society, facilitating the seamless exchange of knowledge between individuals across the globe. However, this connectivity and ease of information dissemination have also accelerated the pace at which humankind acquires new knowledge. Consequently, it is no coincidence that digitization continues to spread rapidly worldwide, gaining increasing importance in private households, businesses, and even (non-)governmental organizations.

1.2 The Omnipresence of Machine Learning and Computer Vision

Machine learning, a field that has had a strong influence on technological advancements, is a noteworthy area of research. Its roots can be traced back to the mid-20th century, but it experienced a significant resurgence in the mid-2010s, driven by remarkable progress in algorithms, particularly those related to (deep) artificial neural networks, and the availability of optimized hardware for accelerated computations.

These algorithms have become the prevailing standard for contemporary computer vision and data processing methods, playing an integral role in numerous products and complex systems. Given the accessibility of knowledge about machine learning techniques and methods, it is reasonable to assume that both governmental and non-governmental organizations will embrace digitization and rely on assistance systems to support them in their endeavors, mirroring the practices adopted by many companies.

Emergency forces, such as the police, fire departments, and rescue services, can particularly benefit from various applications of machine learning. These applications range from forensic analysis to real-time systems that aid in their day-to-day tasks. Considering the prevalence of cameras in our surroundings, whether fixed in public and private spaces or integrated into smartphones carried by individuals, an inconceivable amount of data is being generated. This data holds potential interest for security services when it comes to resolving cases, but it also poses a significant challenge for those responsible for analyzing the vast amount of collected video and image material.

1.3 Perception of Video Surveillance by Society

The perception of video surveillance in European society is a multifaceted topic, considering both its potential benefits and concerns. While it is recognized that video surveillance can enhance public safety and prevent crime, there are valid worries about privacy and civil liberties. One specific area of concern is the use of algorithms, such as face recognition and other biometric evaluation techniques, in video surveillance systems. These algorithms enable behavior analysis and person re-identification, raising questions about invasive monitoring and tracking of individuals. Especially when used by police authorities, the use of behavior analysis in smart video surveillance is a contentious issue. On the one hand, it is argued that such technologies can help identify suspicious activities, detect patterns, and enhance proactive policing. This can potentially contribute to preventing crimes and ensuring public safety. On the other hand, concerns arise regarding the potential misuse of these technologies. There are worries that behavior analyzing algorithms might unfairly target certain individuals or communities, leading to biased outcomes and violations of civil rights. Although citizens of the European Union (EU) acknowledge, that video surveillance has an positive effect on public safety, the possibility of mass surveillance and constant monitoring also raises significant privacy concerns and people call for strict regulatory actions, especially with respect to authorities [Foc17, Spi20]. Different to face recognition, other ways of "smart" systems like those aiming for data privacy show more approval by society as shown in [Gol22a].

In the EU, data privacy is a paramount concern, and strict regulations have been implemented to protect individuals' personal information. The General Data Protection Regulation (GDPR), enforced since 2018, imposes stringent guidelines for collecting, processing, and storing personal data. It requires explicit consent, transparency, and individuals' rights over their data. These regulations have implications for the use of behavior analysis and smart video surveillance by police authorities. They must comply with the GDPR and ensure lawful grounds for data processing. This includes obtaining consent and

providing clear information to the public regarding the purpose and extent of surveillance activities.

However, despite the regulatory framework, concerns regarding privacy and the potential misuse of data persist. Public trust in the responsible use of smart video surveillance technologies by police authorities varies across Europe, where public debates, discussions on ethical implications call for strict oversight and accountability contribute to shaping the perception and acceptance of these technologies.

In conclusion, the perception of video surveillance in European society involves considerations of behavior analysis and the use of smart video surveillance by police authorities. While these technologies offer potential benefits for public safety, concerns about privacy, civil rights, and the potential for bias must be addressed. The GDPR has established regulations to safeguard personal data, but ongoing discussions and calls for transparency and accountability shape public attitudes towards smart video surveillance systems.

1.4 Smart Surveillance from an Ethical Point of View

Smart video surveillance from an ethical perspective raises several important considerations. The literature, such as the renowned work of George Orwell in his novel "1984" serves as a cautionary tale highlighting the potential dangers of unchecked surveillance and its impact on individual freedoms and privacy. Orwell's dystopian vision continues to resonate in discussions about the ethical implications of advanced surveillance technologies.

The situation in autocratic countries, like China, Russia and North Korea provides real-world examples of how intelligent video surveillance can be utilized in ways that raise ethical concerns. Public surveillance and the oppression of citizens are significant aspects that have drawn attention from scholars and decision-makers. The Chinese government e.g., has increasingly utilized digitalization and surveillance technologies to enhance public security and social

control [Ber21b]. This expansion of surveillance has raised concerns about privacy and individual freedoms [Kos21]. The implementation of systems like the Social Credit System (SCS) has been viewed as a tool for surveillance and repression by critics [Xu22a]. The public's perception of surveillance in China is influenced by factors like terrorism concerns, which differ from countries like the United States and Europe [Liu22a]. China's response to public health emergencies, such as the Coronavirus disease 2019 (COVID-19) outbreak, has been a subject of study [Zan20]. The country has made substantial investments in improving its public health system since the 2003 Severe acute respiratory syndrome (SARS) outbreak [Ton15]. Surveillance systems have been crucial in monitoring infectious diseases and responding to outbreaks effectively [Zha21]. However, challenges remain in detecting emerging infectious diseases despite advancements since previous outbreaks like SARS and avian flu [Fen11].

In terms of public security, China has been exporting surveillance technologies, fostering a transnational state-corporate symbiosis to enhance its national security capabilities [Ber21a]. The authoritarian regime in China has been focused on stability maintenance and comprehensive governance, utilizing surveillance technologies as part of its governance agenda [Hua21]. The country's surveillance systems have also been instrumental in managing neglected tropical diseases and parasitic diseases [Lia14, Hao20].

Overall, China's approach to public surveillance reflects a balance between security and control, raising concerns about individual rights and privacy. The country's use of surveillance technologies and systems has implications for public health, security, and governance, shaping the dynamics of citizen-state interactions and societal norms.

The complexity of intelligent video surveillance extends beyond technological aspects. It encompasses sociological, anthropological, and legal considerations. Sociologically, it raises questions about power dynamics, social control, and the potential for discrimination or targeting of specific communities. Anthropologically, it involves analyzing the impact on social norms, behavior patterns, and cultural practices within a surveillance environment. From a

legal standpoint, the use of intelligent video surveillance systems must comply with existing laws and regulations, including data protection and privacy laws. The legality of certain surveillance practices, such as facial recognition, may vary, further complicating the ethical landscape. Ethical discussions surrounding smart video surveillance focus on issues such as consent, transparency, accountability, and the necessity of surveillance measures. Balancing the need for public safety with the protection of individual rights and freedoms is a crucial aspect, of which certain aspects are also addressed by the Artificial Intelligence Act (AI Act) of the European Union [Fri24, Eur24]. The potential for abuse or misuse of intelligent video surveillance technologies cannot be overlooked. Concerns about mass surveillance, constant monitoring, and the diminution of privacy rights require careful consideration and robust safeguards to prevent any infringements on individual liberties.

Engaging in ethical discourse regarding intelligent video surveillance is essential for establishing guidelines, standards, and regulations that prioritize the protection of human rights while leveraging the potential benefits of these technologies. It involves a multidisciplinary approach, incorporating technical expertise, sociological and anthropological insights, legal frameworks, and an ongoing dialogue to navigate the complexities of implementing smart video surveillance systems responsibly.

1.5 From Laboratory to the Street

Methods capable of analyzing collected image and video material hold potential for both forensic analysis and live analysis. However, the task of evaluating such data differs significantly between the two domains. In particular, live analysis presents additional technical constraints, such as the need for faster analysis and limited hardware resources. Moreover, the simultaneous processing of multiple video streams capturing the same situation introduces further challenges, especially a high network traffic coming from those streams, higher computational costs due to the processing of such and many more, which all must be considered during the development of suitable methods.

Figure 1.1: Typical setup at a Public Safety Answering Point (PSAP), where video operators have to work with multiple screens each showing again multiple camera views. [Source: Fraunhofer IOSB]

This thesis focuses on addressing real-world surveillance issues, specifically exploring approaches for analyzing incoming video data. Figure 1.1 gives an example situation where a single operator has to overview many cameras that show various places at the same time. Certain complexities within this subject, including person re-identification and action recognition, have already garnered research interest for several years. Recent advancements, particularly in deep learning algorithms and processing technologies like Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), have propelled the field from predominantly academic to one that is gaining practical relevance for users such as police and fire departments.

Given the specific application field, a myriad of technical, sociological, ethical, and legal challenges arise, further augmenting the complexity of the task at hand. Some examples for technical ones are, but not limited to

- Increasing number of video cameras resulting in higher demands for the available computational resources, and directly affecting the choice of methods and infrastructure,

- Non-cooperative environments, which introduce potentially drastic changes in illumination, harsh contrasts, motion blur, or light flares,
- Challenging single person separation due to many occlusions in crowded scenarios, that may be either dynamic, which cannot be controlled, e.g., during rush hour where people will necessarily occlude each other, or static, e.g., due to structural influences, which in some cases cannot be avoided as well,
- Small pedestrian sizes, due to sensor resolution and the location of the cameras, and potentially large people counts, since these cameras are typically set up yielding overviews over certain areas, and last but not least
- Varying viewing angles, that can be challenging and beneficial at the same time, since occlusions can potentially be diminished, but depending on the type of method applied for analyzing the material, requires additional training data or configurations.

This thesis specifically seeks to research techniques to lessen the amount of information that the analyzing process considers. It strives to be privacy-friendly, which should be accomplished not only during the application phase, or when running on live video streams, but also during the development stage and, consequently, throughout the training of machine learning models. This is primarily driven by two distinct factors: (i) citizen personal rights and informational self-determination; and (ii) ethical issues, such as ethnic background or political stance, which may affect how a data-driven approach would work.

1.6 Aim and Structure of the Thesis

As indicated in the introduction, data privacy is a major challenge for public acceptance of (smart) video surveillance and is therefore the aim addressed in this thesis. Jung [Jun20] examined the topic of privacy-preserving person and human activity recognition and sheds light on various aspects and

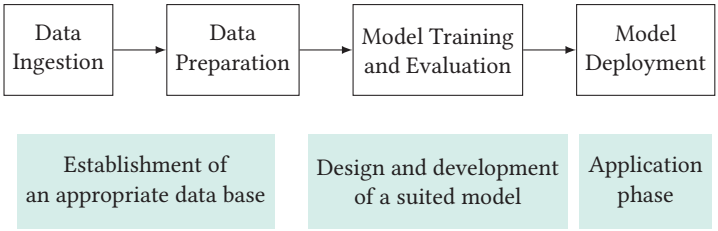


Figure 1.2: Illustration of a typical chain of steps, beginning with the data related operations, the design and training of machine learning models and the final application phase.

vulnerabilities of such systems and methods, also including concerns with respect to data holding and user data. In contrast to Jung [Jun20], this thesis takes a look on the methodical and algorithmic aspects that aim to implement privacy-friendliness. It focuses on how to take privacy-concerns into regard during the overall development of data-driven machine learning approaches. In the field of machine learning, the entire development process consists of different stages at which the prevailing issue can and has to be addressed accordingly. Figure 1.2 depicts the overall procedure.

Each step within the chain offers possibilities to ensure data privacy. However, these steps also introduce challenges that have to be faced.

- **Establishment of an appropriate data base.** It is easy to understand, that in the first place human related data is automatically creating data privacy concerns. In order to design and develop suitable models to apply in surveillance scenarios, it is necessary to have appropriate data with which to train these models. This, however, brings up plenty of challenges. First, such data has to be collected from real-world scenarios, which faces first challenges from a legal perspective. Assuming that these challenges can be solved, the data has to be prepared in order to use it. This includes the need for labels, which can be a very time consuming task especially for data that has to be annotated with fine-grained information like pose information.
- **Design and development of a suited model.** Naturally, the model related steps including the design, but also the development process of

a model, are faced with data and hence exposed to sensitive data. One of the biggest challenges is to design a model that uses only as much data as necessary and to ensure, that especially self-learning models work as intended by the developer. An in-depth understanding about the interplay of the designed model and the used data is inevitable avoid possible privacy issues.

- **Application phase.** Finally, the actual application phase has to cope with the same or at least similar challenges as those already mentioned. At some point, a camera-based approach will have access to full RGB data providing potentially insight in a lot of information about pedestrians. This includes sensitive information, like biometric information or political and/or sexual orientation. Of course, this strongly depends on the chosen approach. Ideally, an approach is chosen, which ensures that as few information about the pedestrians as possible is used. Furthermore, it should be transparent even to non-professionals.

These observations create a frame to the overall roadmap for this dissertation, which is divided into six chapters, starting with the current chapter, i.e., Chapter 1, which introduces the overall topic and field of application of this thesis and motivates the need for this research. Chapter 2 presents the theoretical foundations that are necessary to understand the different approaches and technologies applied within this work. This includes in particular a basic presentation of the field of deep learning and related architectural concepts as well as the introduction of human poses, which are the central human representation used in this work. It is followed by Chapter 3 that gives an overview over existing research that is related to the various topics that are addressed, where Sections 3.1 to 3.3 are mainly related to the topic of building an appropriate data base, and Section 3.4 corresponds to the behavior analysis. Chapter 4 dives into the methodological considerations made to handle the behavior analysis for surveillance setups, which in particular involves topics related to synthetic training data (cf. Sections 4.2 and 4.3), and the human motion analysis (cf. Section 4.4). To be more precise, Section 4.2 addresses the generation of synthetic training data, utilizing a commonly used video game,

in order to generate more realistic synthetic data with the goal to improve the accuracy and robustness of Human Pose Estimation algorithms. The chapter also includes data generation based on a public 3D animation database which is used to enrich the training process of behavior analysis modules as used in this thesis. Furthermore, Section 4.3 tackles a challenge that arises by using synthetic data for Human Pose Estimation, namely the gap between synthetic and real data domain, and therefore focuses on techniques to close the mentioned gap by using generative methods to adapt the real-world domain given the synthetic data. Finally, Section 4.4 delves into the actual behavioral analysis of pedestrians by using the fine-grained structural information provided by an human pose estimator, typically also referred to as *microscopic* or *human-centered* approach. This is compared to another, *macroscopic* or *holistic* method. In Chapter 5 the choices that were made earlier are evaluated and discussed based on different evaluation protocols for the single tasks. Last but not least, Chapter 6 summarizes the contributions and results of this work and takes a look on further possibilities to extend this research.

In summary, the thesis contributes to the fields of Human Pose Estimation, Domain Adaption of synthetic data, and salient behavior recognition, offering novel methodologies and insights that advance the accuracy, adaptability, and interpretability of these areas. The results are then discussed with respect to the overall motivation of a privacy-friendly way of human behavior analysis in the field of video surveillance.

A Note on Implementation

Thomas Golda is responsible for conception and implementation of the overall framework as presented in this thesis. Parts of this dissertation are based on joint works resulting from very close collaboration with his students Andreas Blattmann (Sections 4.2 and 4.3.3), Johanna Thiemich (Section 4.4.3), and David Hohloch (Section 5.3.3). Both, Thomas Golda and the corresponding student, have contributed substantially to this research. While it is difficult to set a precise boundary, Thomas Golda was rather in charge of the idea whereas the student focused on the implementation.

2 Theoretical Background

2.1 Deep Learning

Since various Deep Learning based approaches are used throughout the thesis, this chapter gives a short introduction to artificial neural networks, Deep Learning and the relevant types of neural networks.

2.1.1 Artificial Neuron

Neurons are the basic computational units which compose a neural network architecture. Such a neuron takes an input vector $\mathbf{x} = (x^{(0)}, \dots, x^{(n-1)}) \in \mathbb{R}^n$, and computes an output response $o \in \mathbb{R}$ as shown in Equation (2.1).

$$o := \phi(\mathbf{x}^\intercal \mathbf{w} + b) = \phi\left(\sum_{i=0}^{n-1} x^{(i)} \cdot w^{(i)} + b\right) \tag{2.1}$$

Here, $\mathbf{w} = (w^{(0)}, \dots, w^{(n-1)}) \in \mathbb{R}^n$ is the associated weight vector, $b \in \mathbb{R}$ is its bias and $\phi : \mathbb{R} \to \mathbb{R}$ is the activation function. A typical choice for the activation function ϕ is the Rectified Linear Unit (ReLU), which is a non-linear yet simple and easily differentiable function that is defined as shown in Equation (2.2).

$$\phi(x) := \max(0, x) \tag{2.2}$$

2.1.2 Feedforward Neural Networks

The previously introduced neurons can be stacked both vertically and horizontally, which results in a feedforward neural network. An example for such a network is shown in Figure 2.1.

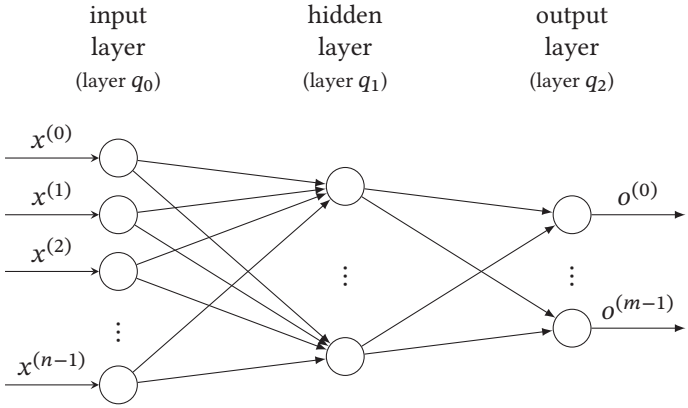


Figure 2.1: Exemplary feedforward network, also called Multilayer Perceptron (MLP). The shown architecture consists of three layers containing multiple neurons each.

Although a single neuron is already capable of simple decision-making tasks, this capability is largely increased when multiple neurons are combined. Neurons stacked vertically form a single layer and stacking those horizontally results in a network, which is considered deep if it comprises multiple layers. Based on this basic structure, a neural network is called fully-connected, if the output of every single neuron in layer q_i is fed into every neuron in layer q_{i+1}. As long as these connections go from layers q_i to q_j with $i < j$ the architecture is refered to as *feedforward* architecture. Finally, the first and the final layer are referred to as the input and output layer, respectively, whereas all other layers are so-called hidden layers.

Overall, given its set of parameters Θ, i.e., its weights and biases, a feedforward neural network defines a mapping $f(\mathbf{x} \mid \Theta) = \hat{\mathbf{y}}$ from the input representation $\mathbf{x} \in \mathbb{R}^n$ to the estimated output representation $\hat{\mathbf{y}} \in \mathbb{R}^m$. It has

been proven [Hor89] that such a network with at least one hidden layer and a non-linear activation function is an universal approximator. This means that it is capable to approximate any continuous function with any desired, nonzero error [Goo16]. Although a network containing a single hidden layer would suffice, the situation is different in practice. Deeper networks, meaning networks with more than just one hidden layer, are mostly used because they typically require far less parameters. This comes from the non-linear relationship of the connections between consecutive layers allowing such networks to generalize better [Goo16].

2.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are neural networks used for processing grid structured input data that has spatial dependencies in local regions [Agg23], e.g., images and graphs. In general, these networks do not differ to those presented earlier, except that the neurons within the layers are arranged in a different way. For a better understanding, CNNs and their differences will be presented in the following.

In a CNN, the input, as well as the neurons in each layer, are organized into 3-dimensional grid structure. The input to each layer q is characterized by its height h_q, width w_q and depth d_q, with $h_q, w_q, d_q \in \mathbb{N}$. For $q > 1$, this input volume is termed to consist of d_q feature maps or channels. Similarly, the neurons of each layer are organized in grid structure as well. The composition of corresponding neurons are called filters. Each filter itself stores weights and is also a 3-dimensional structure, which is again characterized by its height, width and depth. Filters are usually square, meaning that their height and width are equal. Their depth equals the number of channels of the input volume that they are applied to. Generally speaking, the qth layer consists of c_q filters, each of height and width $s_q := h_q = w_q$, and depth d_q, with $c_q, s_q \in \mathbb{N}$. As a result, the output computed from the corresponding input will be a volume containing c_q channels. In addition, each filter is associated with a learnable bias $b_q \in \mathbb{R}$.

CNNs operate in the same manner as feedforward neural networks do, with the difference that the operations of their layers are spatially organized. This operation is called convolution.

2.1.3.1 The Convolution Operation

The traditional convolution operation applies a given filter on every spatial position of the input volume. This is followed by the dot product between this filter, which is flipped in both dimensions, and the subset region of the input. In a CNN, the convolution is performed for all c_q filters within a layer resulting in a 2-dimensional output for each filter. The output volume therefore is 3-dimensional and consists of c_q channels. The traditional convolution operates with a flipped filter, but most deep learning libraries do not flip the filter and implement the convolution with the cross-correlation operation [Goo16]. Although both operations differ by definition, they are equivalent for neural networks since the flipped filters can be learned.

The final resulting convolution operation from layer q to layer $q+1$ can be then defined as

$$h_{ijp}^{(q+1)} \coloneqq \sum_{r=0}^{s_q-1} \sum_{s=0}^{s_q-1} \sum_{k=0}^{d_q-1} w_{(p,q)}^{(r,s,k)} \cdot h_{(q)}^{(i+r\cdot d,\ j+s\cdot d,\ k)} + b_{(p,q)} \tag{2.3}$$
$$\forall i \in \{0, \dots, h_q - s_q\}$$
$$\forall j \in \{0, \dots, w_q - s_q\}$$
$$\forall p \in \{0, \dots, d_{q+1} - 1\}$$

with $\left(w_{(p,q)}^{(i,j,k)}\right) =: \mathbf{W}^{(p,q)} \in \mathbb{R}^{s_q \times s_q \times d_q}$ being the weights, and $b^{(p,q)}$ being the bias of the pth filter in the qth layer [Agg23].

Feature maps in the qth layer are denoted by $\left(h_{(q)}^{(i,j,k)}\right) =: \mathbf{H}^{(q)} \in \mathbb{R}^{h_q \times w_q \times d_q}$. An example of the convolution operation is given in Figure 2.2.

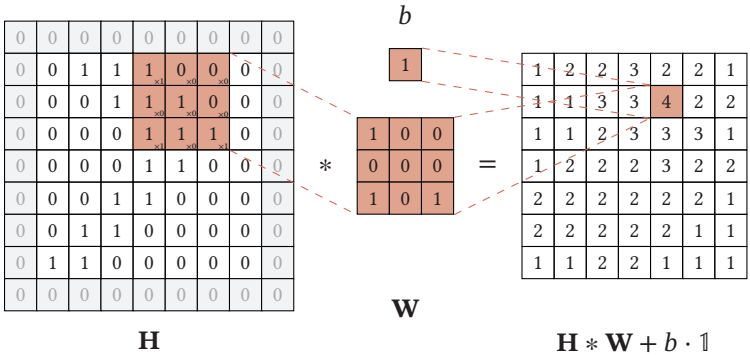


Figure 2.2: Example[a] of the convolution operation as described in Equation (2.3), denoted by ∗. A single channel input $\mathbf{H}$ is convolved with a kernel $\mathbf{W}$ and a bias $b = 1$. The gray shaded parts show the additional padding which is introduced in order to generate an output feature map of same size as the input.

[a] Figure is based on https://tikz.net/conv2d/

Due to the convolution operation, the size of a feature map is reduced after performing the operation. A common way to avoid the reduction in spatial dimensions is padding. Padding is added to the borders of the input, where typically zeros are used since they do not contribute to the output of a convolution.

An important property of the convolution operation is that it is equivariant to translation. This means that if the input was shifted spatially by a certain arbitrary amount, the corresponding values in the feature map would also be shifted accordingly. Hence, certain body features of a person can be equally extracted, irrespective of their location in an image.

2.1.3.2 Graph Convolutional Neural Networks

Since graph-structured data plays an important role throughout this thesis, models that are genuinely designed for such data are also considered. GNNs are designed for the exact purpose. The term Graph Neural Network (GNN) is a general term for NNs that are able to process graph-structured data. Wu

et al. [Wu21] divide GNNs into sub-categories like recurrent GNNs, spatio-temporal GNNs and GCNNs, of which just the latter are relevant to this work.

Classical CNNs as introduced in Section 2.1.3 are usually used for processing images or videos since they have been designed to handle grid-structured data such as pixels in images, different to graph-structured data like human poses. A Graph Convolutional Neural Network (GCNN) contains so-called graph convolutional layers which generalize the traditional neural network convolution operation to graph-structured data. For a traditional 2-dimensional convolution, each pixel in an image can be seen as one node of a graph. As these pixels are arranged in a grid, the actual neighborhood of a single node depends on the size of convolutional filter, resulting in a ordered fixed-sized neighborhood as shown on the left-hand side of Figure 2.3. Here, the weighted average of the neighboring pixels and the center pixel is the result of the application of the convolution filter. A graph convolution however is more complicated since the number of neighbors can differ for each node, which means that the size of the neighborhood is not fixed and the nodes are not ordered. In general, the goal is to calculate the representation of a node v by aggregating both, its own and its neighbors', features [Wu21]. This is shown in Figure 2.3.

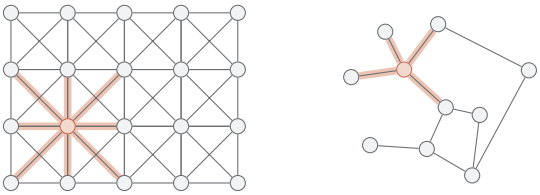


Figure 2.3: Visualization[a] of a traditional 2-dimensional convolution operation on the left and a graph convolution operation on the right. The receptive field of the convolution operation is indicated by the highlighted edges.

[a] Based on a Figure 1 in [Wu21]

Spatial GNNs use information propagation within the graph and are based on the nodes' spatial relations. When processing images, the value in an output feature map of a traditional convolution operation is calculated by summing

up the weighted values of the neighboring pixels. The weights for the summation are defined by the learnable kernel parameters. A spatial graph convolution to a node v, yields a new representation by combining the features $\mathbf{x}_u$ of its neighboring nodes $u \in \mathcal{N}(v)$ and its own features $\mathbf{x}_v$ in an overall feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d_x}$. Here, $\mathcal{N} : V \to \mathcal{P}(V)$ is the neighborhood function, $d_x \in \mathbb{N}$ is the dimension of the input features $\mathbf{x}_i$ with $i \in \{v\} \cup \mathcal{N}(u)$ and where $n = |\{v\} \cup \mathcal{N}(u)| \in \mathbb{N}$. Stacking multiple graph convolutional layers with each layer having its own set of learnable weights results in a high-level representation of the nodes within the graph. [Wu21]

According to Chen et al. [Che20], this spatial graph convolution process can mathematically be expressed as:

$$\mathbf{H}^{(l+1)} = f(\, \hat{\mathbf{D}}^{-1} \cdot \hat{\mathbf{A}} \cdot \mathbf{H}^{(l)} \cdot \mathbf{W}^{(l)} \,) \tag{2.4}$$

Here, $\mathbf{H}^{(l)} \in \mathbb{R}^{n \times d_l}$ represents the feature matrix of the lth layer which has n nodes and d_l features. The first and the last layer's features are $\mathbf{H}^0 =: \mathbf{X}$ and $\mathbf{H}^L =: \mathbf{Z}$ respectively. $\hat{\mathbf{A}}$ refers to the adjacency matrix and $\hat{\mathbf{D}}$ is the diagonal node degree matrix of $\hat{\mathbf{A}}$ therefore determines the way in which node features are propagated to neighboring nodes. $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$ is a trainable weight matrix mapping the d_l-dimensional features into d_{l+1}-dimensional features. The function f usually defines a non-linear activation function. [Che20]

Spatial graph convolutions do not consider spectral behavior [Che20]. Zhang et al. [Zha19] showed that spatial Graph Convolutional layers are low-pass filters and therefore might not be able to extract meaningful features in high frequency areas. But overall, they are much more commonly used than spectral graph convolutions due to several reasons as explained in [Wu21]. Spatial models are more efficient and scalable to large graphs since the convolution operation can directly be performed in the graph domain and it can be executed in batches of nodes. Furthermore, spatial convolution operations can be performed on both directed and undirected graphs while spectral convolution operations can be applied on undirected graphs only. Lastly, models using spatial convolution operations generalize better to changing graphs. These properties influenced the choice of spatial GCNNs for this thesis.

For further information about spatial graph convolutions as well as the related spectral graph convolutions see [Che20, Wu21] since this would exceed the scope of this thesis.

2.1.3.3 Fully Convolutional Neural Networks

CNNs are powerful models when it comes to classifying entire images. But there are also other problems where a classification at pixel level is required, e.g., as it is the case with semantic segmentation. Fortunately, only two adjustments need to be made in order for CNNs to cope with this task. The first one is to replace the fully-connected layers with convolutional layers, where the number of channels equals the size of the replaced layer. It is common to use 1×1 kernels here and with this, a convolution can be regarded as applying the previously replaced fully-connected layer to every spatial location of the feature map. Since only convolutional layers are present in this architecture, such networks are termed fully convolutional [She17]. This also means that these networks can process images of arbitrary size.

However, with this adaptation only, the output is still smaller than the input size due to pooling. Therefore, the second adjustment is to introduce a final upsampling layer that will scale the estimated output such that it matches the input size. In its simplest form, the upsampling is fixed, e.g., it is implemented with bilinear interpolation, but it can also be learned.

2.1.3.4 Autoencoder

The Autoencoder (AE) is a specific deep neural network architecture that aims to find a low-dimensional representation of data, which still holds the relevant information. An AE consists of two main parts: the encoder and the decoder. The encoder subnetwork compresses the input data into a lower-dimensional latent representation, often referred to as the bottleneck layer. In an opposite manner, the decoder subnetwork attempts to reconstruct the original input

data from this compressed representation. The aim is to produce a reconstruction that is as close as possible to the original input, effectively learning a compact and meaningful representation of the data. Figure 2.4 shows a schematic of an AE-architecture.

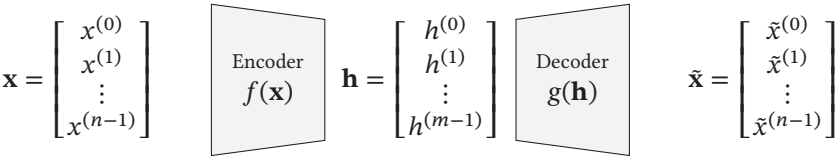


Figure 2.4: Schematic architecture of an AE. It is characterized by the bottleneck $\mathbf{h}$ between the encoder and decoder part.

As already stated, the network is trained to reproduce the initial input $\mathbf{x}$ from the resulting embedding $\mathbf{h}$. The reconstruction is done by the decoder and yields the reconstructed input $\tilde{\mathbf{x}}$. In order to do so, the encoder learns a mapping function $f(\mathbf{x}) = \mathbf{h}$ which maps the high-dimensional input $\mathbf{x} \in \mathbb{R}^n$ into a usually lower-dimensional representation $\mathbf{h} \in \mathbb{R}^m$ with $m < n$. The decoder takes the embedding $\mathbf{h}$ as input and learns a function $g(\mathbf{h}) = \tilde{\mathbf{x}}$ in order to return a reconstruction $\tilde{\mathbf{x}}$ of the input $\mathbf{x}$ [Goo16].

The most interesting part of this architecture for most applications is the embedding $\mathbf{h}$, since it should hold the most important information. When the hidden dimension is smaller than the input dimension, the model is more specifically called an undercomplete AE [Goo16]. To achieve this, an AE is usually trained by minimizing the reconstruction error, which can be e.g., the mean squared error between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ [Ala14].

Since the concept of AEs is quite general, it can be potentially used with any kind of neural network. For processing image data, AEs are built using convolutional layers in both, encoder and decoder [Mar20]. On graph-structured data, graph convolutional layers can be used [Gon19]. Hence, both count as fully-convolutional architectures as introduced in Section 2.1.3.3. Despite these choices, standard fully-connected layers are also suitable to use within the encoder and decoder component.

Since AEs are trained to reconstruct the original input, they can be trained in an unsupervised manner, where no labels for the samples are available. This makes them versatile for many different applications.

2.1.4 Recurrent Neural Networks

The so far presented networks are only able to process single static inputs. In general, one can easily imagine problems where sequences of data samples need to be processed. Such sequences provide additional information that might exhibit useful or even relevant patterns that are necessary for the overall decision making process. One type of architecture that is capable of doing so are so-called recurrent neural networks.

The most important property of recurrent neural networks is the additional feedback loop that is added at every hidden neuron, compared to the earlier introduced feedforward neural networks. This way, the hidden state from the previous step is also considered during computation. A major drawback of these networks is, that they can easily suffer from the so called vanishing gradient problem [Hoc98], which can prevent such networks from learning.

Today, the most widely used types of these networks that aim to overcome the vanishing gradient problem, are the Gated Recurrent Unit (GRU) [Cho14] and Long Short-Term Memory (LSTM) [Hoc97, Ger00], where recurrent neurons are replaced by GRU- or LSTM-cells respectively.

2.1.4.1 Long Short-Term Memory

An LSTM-cell stores a state vector $\mathbf{c}_t$, namely cell state, for every timestep $t \in \{0, \dots, T-1\}$, where T is the length of the input sequence. Its output is the hidden state $\mathbf{h}_t$, which is a non-linear transform of the current cell state.

The latter is obtained by manipulating $\mathbf{c}_{t-1}$ using three gates, namely input-gate $\mathbf{i}_t$, output-gate $\mathbf{o}_t$ and forget-gate $\mathbf{f}_t$. Based on the current input $\mathbf{x}_t \in \mathbb{R}^n$ and $\mathbf{h}_{t-1}$, $\mathbf{i}_t$ and $\mathbf{f}_t$ determine the amount to reuse from $\mathbf{c}_{t-1}$ and the amount to incorporate from both $\mathbf{x}_t$ and $\mathbf{h}_{t-1}$. Finally, $\mathbf{o}_t$ determines the amount

to output from $\mathbf{h}_t$. All of the aforementioned vectors are of same size, i.e., $\mathbf{c}_t, \mathbf{h}_t, \mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t \in \mathbb{R}^m$.

In summary, the LSTM cell can be described as

$$\mathbf{i}_t \coloneqq \sigma\left(\mathbf{W}_{xi} \cdot \mathbf{x}_t + \mathbf{W}_{hi} \cdot \mathbf{h}_{t-1} + \mathbf{b}_i\right) \tag{2.5}$$

$$\mathbf{f}_t \coloneqq \sigma\left(\mathbf{W}_{xf} \cdot \mathbf{x}_t + \mathbf{W}_{hf} \cdot \mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{2.6}$$

$$\mathbf{c}_t \coloneqq \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \phi\left(\mathbf{W}_{xc} \cdot \mathbf{x}_t + \mathbf{W}_{hc} \cdot \mathbf{h}_{t-1} + \mathbf{b}_c\right) \tag{2.7}$$

$$\mathbf{o}_t \coloneqq \sigma\left(\mathbf{W}_{xo} \cdot \mathbf{x}_t + \mathbf{W}_{ho} \cdot \mathbf{h}_{t-1} + \mathbf{b}_o\right) \tag{2.8}$$

$$\mathbf{h}_t \coloneqq \mathbf{o}_t \odot \phi\left(\mathbf{c}_t\right) \tag{2.9}$$

with $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xc}, \mathbf{W}_{xo} \in \mathbb{R}^{m \times n}$, $\mathbf{W}_{hi}, \mathbf{W}_{hf}, \mathbf{W}_{hc}, \mathbf{W}_{ho} \in \mathbb{R}^{m \times m}$ and $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_c, \mathbf{b}_o \in \mathbb{R}^m$, resulting in $4 \cdot (mn + m^2 + m)$ learnable parameters overall. The element-wise product is denoted by $\odot$, whereas ϕ and σ are non-linear activation functions.

A common activation function ϕ is the hyperbolic tangent (tanh), while σ is fixed to be the sigmoid function. Both are defined as

$$\tanh(x) \coloneqq \frac{e^{2x} - 1}{e^{2x} + 1} \in (-1,1) \tag{2.10}$$

and

$$\sigma(x) \coloneqq \frac{1}{1 + e^{-x}} \in (0,1) \tag{2.11}$$

It is easy to understand that these functions enable the previously introduced interpretation of gates and cell state. Note that plugging in a vector into these functions means that they are applied element-wise.

2.1.4.2 Gated Recurrent Units

GRUs are similar to LSTMs, but with a reduced number of gates. The input, output- and forget-gate of LSTMs are replaced by a reset-gate $\mathbf{r}_t$ and an

update-gate $\mathbf{u}_t$ for GRUs.

$$\mathbf{r}_t \coloneqq \sigma\left(\mathbf{W}_{xr} \cdot \mathbf{x}_t + \mathbf{W}_{hr} \cdot \mathbf{h}_{t-1} + \mathbf{b}_r\right) \tag{2.12}$$

$$\mathbf{u}_t \coloneqq \sigma\left(\mathbf{W}_{xu} \cdot \mathbf{x}_t + \mathbf{W}_{hu} \cdot \mathbf{h}_{t-1} + \mathbf{b}_u\right) \tag{2.13}$$

$$\mathbf{h}_t \coloneqq (1 - \mathbf{u}_t) \odot \phi(\mathbf{r}_t \odot \mathbf{W}_h \cdot \mathbf{h}_{t-1} + \mathbf{W}_x \cdot \mathbf{x}_t + \mathbf{b}) + \mathbf{u}_t \odot \mathbf{h}_{t-1} \tag{2.14}$$

where $\mathbf{W}_{xr}, \mathbf{W}_{xu}, \mathbf{W}_x \in \mathbb{R}^{m \times n}$, $\mathbf{W}_{hr}, \mathbf{W}_{hu}, \mathbf{W}_h \in \mathbb{R}^{m \times m}$ and $\mathbf{b}_r, \mathbf{b}_u, \mathbf{b} \in \mathbb{R}^m$ are the learnable parameters. The element-wise product is denoted by $\odot$, whereas ϕ and σ are non-linear activation functions analogous to those introduced for LSTMs. In total, GRUs have $3 \cdot (mn + m^2 + m)$ learnable parameters and are hence just 75% of the size of LSTMs [Cho14, Mat21].

2.1.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have been introduced in 2014 and have since become a widely investigated field and currently one of the most popular form of generative models. Despite the fact that they are not explicitly yielding the desired probability distribution, they have been successfully applied to many fields of computer vision such as image synthesis [Kar21] and image-to-image-translation [Zhu17]. In the following description of the GAN framework, it is referred to the classical formulation of this type of neural networks [Goo14], which is called VanillaGAN.

The GAN setup introduces two neural networks, called *generator* and *discriminator* respectively, in addition to a given dataset consisting of samples $\mathbf{x}_r$ from an unknown probability distribution $p(\boldsymbol{X}_r)$. The generator is a generative model that tries to mimic this distribution given input signals $\mathbf{z}$, which are sampled from the prior distribution $p(\boldsymbol{Z})$. Thus, the generator is actually just a mapping $G(\mathbf{z}; \Theta_G) = \mathbf{x}_G$ from this prior to the target distribution $p(\boldsymbol{X}_r)$, where Θ_G denotes the learnable parameters of the generator network. The discriminator network $D(\mathbf{x}; \Theta_D)$ on the other side is a binary classifier that takes as input image samples $\mathbf{x}_r$ and $\mathbf{x}_G$ from both target data and generated data and outputs a scalar value, which represents the probability of a sample

belonging to the target dataset. During the optimization process, these networks are optimized against each other in the following way: The discriminator updates its learnable parameters Θ_D such that it is able to perfectly differentiate between data originating from the target distribution and data generated by the generator network, which, for its part, updates its parameters Θ_G such that the generated data exactly meets the distribution $p(\boldsymbol{X}_r)$. Hence, both networks play the following two-player minimax game with the loss function $\mathcal{L}(G,D)$:

$$\begin{aligned}\min_G \max_D \mathcal{L}_G(G,D) = \mathbb{E}_{\mathbf{x}_r \sim p(\boldsymbol{X}_r)}[\log D(\mathbf{x}_r)] \\ + \mathbb{E}_{\mathbf{z} \sim p(\boldsymbol{Z})}[1 - \log D(G(\mathbf{z}))]\end{aligned} \tag{2.15}$$

Goodfellow et al. [Goo14] showed that the optimal value of the objective function as displayed in Equation (2.15) is defined by a generator that meets the target distribution $p(\boldsymbol{X}_r)$ and a discriminator that outputs the probability of 0.5 for all samples that are classified. They state that from a game theoretical point of view, this state is the nash equilibrium [Nas51] between the generator and the discriminator.

2.1.5.1 Training and Convergence

The training process of GANs is a challenging task. Although, [Goo14] showed the theoretical convergence for the GAN framework, the training suffers from the non-convex and high-dimensional nature of the GAN-game. Furthermore, real-world data distributions usually show a higher complexity and multimodality, as it is the case for image processing [Sal16]. An often-faced issue, known as *mode collapse*, which avoids the generator G to capture the desired distribution, arises in the following scenario:

(Step 1) G captures only a small number of modes of the target distribution what causes D to categorize these modes as coming from the generator and samples lying within the other modes as belonging to the real data

(Step 2) G exploits D by switching modes to the ones categorized as being modes of the real-world data before

(Step 3) D categorizes samples lying within the modes which G had captured initially as belonging to the real data

(Step 4) Return to (Step 1)

Instead of stabilizing and converging to an equilibrium, a GAN framework that faces this issue begins to oscillate as stated above. Throughout the last years, many so-called *regularization* strategies have been proposed to overcome this issue. Most of them only are only applicable to certain scenarios while negatively affecting other training properties as training speed. The majority of these regularization methods aims limiting the power of the discriminator during the training process.

In [Mes18] the authors show that even when building generator and discriminator from simple linear models the classical GAN-objective as shown in Equation (2.15) forces D to push G away from the equilibrium, when the generated distribution arrives at this point. This behavior is identified as one potential reason for the oscillating behavior. Based on this observation, the following regularization term has been derived

$$\min_D R = \gamma \cdot \mathbb{E}_{\mathbf{x}_r \sim p(\boldsymbol{X}_r)}[\|\nabla_{\Theta_D} D(\mathbf{x}_r)\|_2^2] \tag{2.16}$$

which is denoted *one-sided gradient penalty* and added to the optimization objective that is depicted in Equation (2.15). By penalizing large gradients of D solely for examples coming from the real-world data distribution, the regularization term given in Equation (2.16) introduces a better stability of the training procedure for an appropriately chosen value of the weighting factor γ. This term is included by default to all the GANs that are examined within this thesis unless stated otherwise.

2.1.6 Transformer

This subsection presents the Transformer model, introduced by Vaswani et al. [Vas23] in 2017, which is the first sequence transduction model based entirely on an attention mechanism, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention. Transformer has emerged as a prominent framework within the computer vision community, showcasing remarkable potential in sequence modeling [Yu24]. Initially presented for the task of Natural Language Processing (NLP), it has demonstrated competitive performance in comparison to CNN-based approaches across various tasks such as image classification [Dos20, Liu21b, Tou21], object detection [Car20], Human Pose Estimation (HPE) [Xu22b, Xu24] and semantic segmentation [Zhe21].

Vaswani et al. [Vas23] state that competitive neural sequence transduction models typically use an encoder-decoder structure. Given $\mathbf{z}$, the decoder then generates an output sequence $\{\mathbf{y}_0, \ldots, \mathbf{y}_{m-1}\}$ one token at a time, following an auto-regressive approach [Gra13]. The Transformer implements this architecture with stacked self-attention and fully connected layers in both the encoder and decoder.

2.1.6.1 Encoder

The encoder part of the Transformer is constructed with one or multiple layers, each layer comprising two sub-layers. The first sub-layer utilizes a multi-head self-attention mechanism, while the second employs a position-wise fully connected feed-forward network. Residual connections [He16] surround each sub-layer, followed by layer normalization [Ba16].

2.1.6.2 Decoder

Similar to the encoder, the decoder consists of at least one layer. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer in between, which performs multi-head attention over the output of

the encoder stack. Residual connections and layer normalization are applied around each sub-layer. The self-attention sub-layer in the decoder stack is modified to prevent positions from attending to subsequent positions. This masking, combined with the fact that the output embeddings are offset by one position, ensures that predictions for position i can depend only on the known outputs at positions j with $j < i$.

2.1.6.3 Positional Encoding

In the absence of recurrence and convolution in the model, incorporating information about the sequence's order is crucial. To achieve this, *positional encodings* are introduced at the input of the encoder and decoder stacks. Therefore, ensuring compatibility by summation with the input embeddings. These positional encodings, having the same dimension $\sqrt{d_{model}}$ as the embeddings, come in various types, including both learned and fixed options.

2.1.6.4 Vision Transformer

Transformers have been successfully applied to other fields than NLP, which includes the field of computer vision. The Vision Transformer (ViT) [Dos20] is one prominent example designed for image classification tasks by directly processing sequences of image patches and tries to follow the architecture of the transformer model as much as possible. Therefore, ViT splits the input image $\mathbf{J} \in \mathbb{R}^{h \times w \times c}$ into a sequence of 2-dimensional patches of size $p \times p \times c$ and generates a flattened representation $\mathbf{x} \in \mathbb{R}^{p \cdot p \cdot c}$ where c is the number of channels and h and w represent the resolutions of the original image. Consequently, the effective sequence length for the transformer is calculated as $n = \frac{h \cdot w}{p^2}$. Utilizing constant widths across layers, a trainable linear projection maps each vectorized patch to the model dimension d_{model}, producing patch embeddings. Similar to BERT's [CLS] ("class") token [Dev18], a learnable embedding is applied to the sequence of embeddings, serving as the image representation. During both pre-training and fine-tuning stages, classification heads are consistently attached. Additionally, 1-dimensional

position embeddings are added to the patch embeddings for retaining positional information. Notably, ViT employs only the standard transformer's encoder (except for layer normalization), with its output preceding an MLP head. ViT is typically pre-trained on large datasets and fine-tuned for downstream tasks with smaller data. ViT performs well when pre-trained on large datasets, from 14 million up to 300 million images, surpassing inductive bias limitations. Despite modest results on datasets with less data like ImageNet, ViT achieves state-of-the-art performance on image recognition benchmarks when pre-trained at sufficient scale, such as the JFT-300M dataset [Sun17].

2.2 Human Poses

This section introduces the underlaying representation of humans for this thesis. Therefore, Section 2.2.1 introduces a basic definition of the fundamental feature representation of pedestrians for this thesis, followed by Section 2.2.2 which shortly presents other existing human body models. If not stated differently, throughout this thesis a human pose refers to the skeleton representation as introduced in the following.

2.2.1 Definition

A human pose is an abstract representation of a single person for a given timestep t. Each pose contains multiple keypoints that correspond to body joints or special body parts like, for instance eyes, ears and nose. More generally speaking, a pose is an undirected graph $G_t = (V_t, E)$, where a vertex $\mathbf{v} = (v^{(0)},v^{(1)},v^{(2)}) \in V_t \subset \mathbb{R}^2 \times \{0,1,2\}$ corresponds to a certain keypoint. These keypoints consist of three components, where $v^{(0)}$ and $v^{(1)}$ are the x- and y-coordinates and $v^{(2)}$ is the visibility of the corresponding keypoint. An edge $e \in E \subset V_t \times V_t$ connects somatically or semantically related keypoints. Apart from this graph representation, such poses are also represented as a

matrix, where each row corresponds to keypoint

$$\mathbf{V}_t \coloneqq \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_{K-1} \end{bmatrix} \in \mathbb{R}^{K \times 3} \tag{2.17}$$

In current state-of-the-art methods a pose G_t is represented as a set of heatmaps $H = \{ \mathbf{H}_k \in \mathbb{R}^{w \times h} \mid k \in \{0, ..., K-1\} \}$, where the value at the position $\mathbf{H}_k^{(x,y)}$ denotes a per-pixel likelihood for keypoint k. In order to obtain the position of $\mathbf{v}_k$ the maximum of the corresponding heatmap $\mathbf{H}_k$ has to be determined.

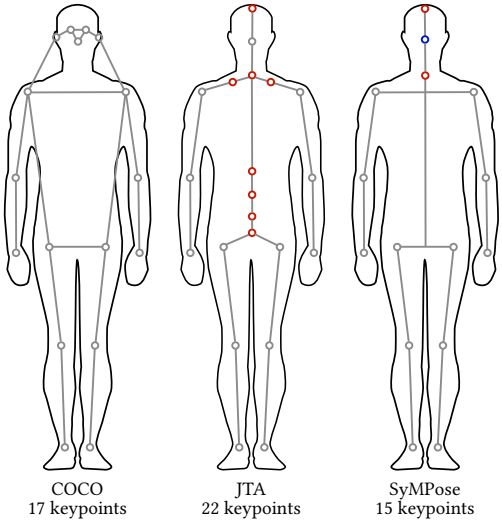


Figure 2.5: Schematical visualization of a human pose for three exemplary models. Depending on the selected model, the types and number of keypoints may vary. The models differ mostly with respect to the torso and head part, and the level of detail for hands, whereas the limbs are usually represented in the same way.

Apart of the general definition of a human pose, there are various models [And14, Lin15, And18, Fab18] that differ in the choice of keypoints, as

well as their connections. Each of these models were introduced for a different reason and therefore provide more keypoints for body parts with more focus on, e.g., face or hands. The main goals of using human poses are fast to process and easily comprehensible features, that reduce a person to its basic physical structure. The reduced information should furthermore facilitate non-discriminatory behavior analysis, since for instance information on skin color, ethnicity, or political and sexual orientation are omitted. Figure 2.5 shows some examples of different pose models.

2.2.2 Human Body Models for Pose Estimation

As indicated at the beginning of the current sections, 2-dimensional keypoint representations ("stick-figures") are just one possible representation for human poses. Dubey et al. [Dub22] summarize the field of human body models into three different categories, namely

- *kinematic* or *skeleton-based* models,
- *planar*, and
- *volumetric*.

Kinematic models reflect human body structure in a comprehensible way that comprises a variety of joint configurations and limb orientations either 2- or 3-dimensionally. As a result, this model is utilized to represent the relationships between various body parts. The 3-dimensional case offers additional information in spatial space and removes ambiguities that can arise when working with 2-dimensional (i.e., flat) information [Sig09, Ion14]. Although there has already research been done [Wei16a, Tom17, Mar17, Isk19] on creating such, in the context of video surveillance they are difficult to obtain. Sensors, delivering enough data to extract the needed information (e.g., depth sensors, stereo cameras) are often not desired by the authorities. Another way to get the lacking information could be multi-camera setups, that can be used in order to get the additional information, in practice these cameras are often mounted to far away from each other, which will yield a poor pose estimation performance.

Planar models, also referred to as a contour-based models, are primarily used to represent 2-dimensional body contours. In contrast to the kinematic model, they express the human body's appearance and shape. Usually, body parts are represented by multiple rectangles approximating the human body contours [Bra14].

Finally, volumetric models, which are essentially representations in a 3-dimensional space, encompass human body shapes and poses represented by volume-based models with geometric shapes or meshes. Earlier geometric shapes for modeling body parts include cylinders or conics [Sid00], whereas modern volume-based models are represented as meshes, normally captured using 3-dimensional scans. Widely used volume-based models includes Shape Completion and Animation of People (SCAPE) [Ang05], Skinned Multi-Person Linear Model (SMPL) [Lop15], and a unified deformation model [Joo18]. Recently dense pose estimation [Gül18] has gained an increasing amount of interest from the research community [Nev19, Dua21, Rak21]. Dense human pose estimation aims at mapping all human pixels of an RGB image to the 3-dimensional surface of the human body. Although, researchers have achieved promising results using this technique, they are still limited to rather high resolution imagery and can be seen as they are still in their infancy. For further details on the mentioned body models see [Dub22].

3 Related Work

3.1 Human Pose Estimation

This thesis uses skeletal information as a central feature for performing privacy-friendly behavior analysis. Therefore the following sections give an overview over existing methods for HPE and ways to obtain human skeletal representations, followed by an overview over publicly available datasets to train such methods. These datasets are separated into real-world and synthetic ones.

3.1.1 Methods

As stated in Section 2.2.2, Dubey et al. [Dub22] provide a categorization of various human body models. They also state, that there are various other ways to classify existing approaches for HPE, like the type of input modality, the number of cameras, the type of input data and the number of tracked people. With the setting of public safety and video surveillance in mind, this section takes a look from two different perspectives. Firstly, by addressing so-called Multi-Person Pose Estimation (MPPE) setup, where subject to the overall HPE process are multiple people, which stands in contrast to the Single-Person Pose Estimation (SPPE). The former represents the typical situation for Closed-Circuit Television (CCTV). Secondly, this sections looks at different input-modalities.

Methods for MPPE are typically separated in two main paradigms: *top-down* and *bottom-up*. Top-down approaches apply pre-existing person detectors, like e.g., You Only Look Once (YOLO) introduced by Redmon et al. [Red16],

to identify individual people in input images. Then, SPPE is applied to each person's bounding box to create poses for all pedestrians. Unlike top-down approaches, bottom-up methods identify all body joints in a single image first, before associating them to separate subjects. The number of individuals in the input image will have a direct impact on the processing time in the top-down pipeline. Bottom-up methods often have a faster computing speed than top-down methods since they do not require individual posture detection for each user [Dub22, Zhe23]. Some of the most prominent and widely used methods for HPE are AlphaPose [Fan23], OpenPose [Cao21], HRNet [Sun19], and ViTPose [Xu22b]. AlphaPose, as described by Fang et al. [Fan23], is a top-down approach designed for whole-body regional multi-person pose estimation and tracking in real-time. This method excels in accurately localizing whole-body keypoints and tracking humans simultaneously, even when provided with inaccurate bounding boxes and redundant detections. By first detecting human bounding boxes and then estimating poses within each box independently, AlphaPose demonstrates efficiency in real-time applications. On the other hand, OpenPose, introduced by Cao et al. [Cao21] is a bottom-up approach for multi-person 2-dimensional pose estimation. This method utilizes part affinity fields to associate detected keypoints of body parts into complete poses. As hardware capabilities improve, bottom-up methods like OpenPose can leverage higher resolutions to potentially reduce the accuracy gap compared to top-down approaches. OpenPose's focus on detecting keypoints and forming complete poses offers a different perspective in the realm of pose estimation methodologies. Sun et al. [Sun19] proposed a method called High-Resolution Net (HRNet) for HPE that emphasizes deep high-resolution representation learning. By recovering high-resolution representations from low-resolution ones generated by a high-to-low resolution network, this approach aims to enhance the quality of pose estimation results. This method falls under the top-down paradigm by detecting person instances using a person detector and then predicting detection keypoints. Finally, ViTPose, introduced by Xu et al. [Xu22b], takes a unique approach by employing a Vision Transformer-based method for HPE. By utilizing a plain and non-hierarchical ViT along with simple deconvolution decoders, ViTPose aims to provide a baseline for pose

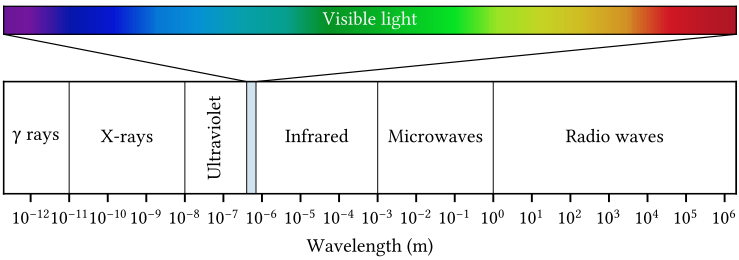


Figure 3.1: Overview of the electromagnetic spectrum based on information from the National Aeronautics and Space Administration [Nat10]. Most methods for HPE fall into the extended range of visible light including the infrared (RGB+IR), where NIR typically lays in a range of 0.75-1.4µm (active), and TIR in a range between 8-15µm (passive) [DAm09].

estimation tasks. This method represents a novel direction in pose estimation research, exploring the capabilities of ViT in this domain.

In summary, each of these approaches - AlphaPose, OpenPose, HRNet, and ViTPose - contributes distinct methodologies to HPE, either through top-down or bottom-up strategies.

As mentioned earlier, the second perspective takes a look on the input modalities. When talking about HPE, literature typically refers to the case of data collected using RGB-sensors, either as single images or as videos. However, the research community has also reported achievements in extracting human pose information using other sensors, which in essence can possibly eliminate the privacy-related drawbacks of RGB data. In the following, this section gives some examples for approaches that do not solely rely on RGB data, and hence can be associated to different ranges of the electromagnetic spectrum, as given in Figure 3.1. To be more specific, existing methods for HPE can be found in the range from around $5 \cdot 10^{-7}$ meters, i.e., the lower end of the visible spectrum, up to radio waves around 10^{-1} meters.

Starting with the infrared, i.e., near infrared (NIR) and thermal infrared (TIR), different techniques follow different approaches, to either generate an 2-dimensional image or directly obtain a 3-dimensional point cloud. Zhu et al. [Zhu22] propose a dual-channel network for 3-dimensional HPE, where

one channel applies a classical pre-trained HPE-algorithm, namely OpenPose [Cao21], which has been presented earlier, and the other channel uses a Transformer-based architecture to derive depth-information from the thermal infrared image sequence. Guo et al. [Guo22] on the other hand propose a novel benchmark to assess the identity-preserving HPE-task by establishing three baseline methods based on different YOLO detectors. These are accompanied with a manually annotated dataset, called Identity-Preserved Human Posture Detection in Thermal Images (IPHPDT), which is based on a previously published thermal dataset and extends it with posture annotation. The baseline models are extended with an additional posture prediction head. Similar to Guo et al. [Guo22], Smith et al. [Smi23] introduce a novel thermal image database. The authors investigate eight state-of-the-art HPE-methods in the classical manner as known from the Common Objects in Context (COCO) dataset. They tackle the lack of thermal image data to fully train models, by converting existing datasets to grayscale images and fine-tuning on their proposed UCH-Thermal-Pose database, and showed that almost all investigated methods benefit from the grayscale transformation, with ViTPose achieving the best results. Finally, Lupión et al. [Lup24] propose the use of several cameras following a multi-view approach, increasing the ability of their solution to infer 3-dimensional poses and avoiding occlusions. Each thermal camera is part of an IoT device that consists of a classical RGB camera and a TIR sensor. On top of the intrinsic and extrinsic parameters for both cameras, the authors propose to use a homography to obtain fully paired images. They furthermore propose a novel method called ThermalYOLO, to fill the gap between the thermal and visible domains in terms of people detection. This method provides bounding boxes in the visual part of the sensor, which is then mapped using the determined homography to obtain annotations in the thermal image, which, in turn, is used to fine-tune YOLO on thermal images. From the overall camera setup they then generate the desired 3-dimensional pose information.

LiDAR-based HPE is a technique that utilizes Light Imaging, Detection and Ranging (LiDAR) technology to estimate the 3-dimensional pose of a human body. This process involves analyzing LiDAR data, which provides detailed

spatial information about the environment by emitting laser pulses and measuring the time it takes for them to return after hitting objects. By integrating LiDAR data with other modalities such as RGB images or full motion video, it is possible to accurately estimate the pose of a human subject in 3-dimensional space. The following papers [Gla19, Fur21, Zan23] address the problem of HPE using LiDAR technology from different perspectives: Glandon et al. [Gla19] proposed a method for 3-dimensional HPE and human identity recognition using LiDAR full motion video. This work not only estimates human poses in 3-dimensional space but also addresses the challenge of identifying individuals based on their skeletal structure using LiDAR full motion video data. The first step is to determine the range information to focus on processing, which is done by extracting a silhouette from the raw LiDAR video. Using this silhouette as a seed, and guided extracting walk direction, their algorithm estimates 13 3-dimensional joint positions. The authors examine the performance of their proposed method using an internal dataset consisting of ten different subjects, on which they compare results with the OpenPose pose estimator. Furst et al. [Fur21] introduced the 3D Human Pose Estimation from RGB and LiDAR (HPERL) method, which focuses on combining RGB and LiDAR data for accurate HPE. Their architecture processes the RGB images and LiDAR-generated point clouds as input modalities, using an RPN as feature extraction stage, followed by pose estimation stage predicting scores and deltas for K anchor poses. In contrast to other approaches, the anchor poses are generated from the 3-dimensional boxes of the first stage. By adding the deltas to these anchors and selecting based on the classification scores, the poses are predicted. This approach aims to enhance the robustness and accuracy of pose estimation systems by integrating information from both modalities. Finally, Zanfir et al. [Zan23] presented the Semi-supervised Multi-modal 3D Human Pose Estimation for Autonomous Driving (HUM3DIL) method, which targets HPE in the context of autonomous driving. This work explores how LiDAR data, combined with other modalities, can be used to estimate human poses in real-world scenarios relevant to autonomous vehicles. The semi-supervised aspect suggests leveraging both labeled and unlabeled data to improve pose estimation accuracy.

In conclusion, LiDAR-based HPE involves using LiDAR technology along with other modalities to estimate a 3-dimensional pose of humans. All presented work offers unique perspectives and methodologies to address this problem, showcasing the diverse applications and advancements in the field of HPE using LiDAR technology.

Leaving the range of classical visual and infrared range, shows that longer wavelengths have been successfully utilized for HPE as well. For instance, Zhao et al. [Zha18b] extracted human poses using radio signals. To do so, they train a teacher-student network for cross-modal supervision and utilize the fact, that the human body is specular in the frequency range that traverse walls. The teacher, generating human pose annotation from the visual domain generates keypoint heatmaps that are used as target values for the student network. With respect to performance, the authors claim that their approach achieves competitive results and is less likely to generate false positives where the visual model is likely to generate such. The major drawbacks mentioned by Zhao et al. [Zha18b] are the low spatial resolution coming from the physical setup, and the limited operating distance which is dependent on its transmission power. Similarly, Geng et al. [Gen23] developed a deep neural network that maps the phase and amplitude of WiFi signals to UV coordinates within 24 human regions. The results of the study reveal that their model can estimate dense poses of multiple subjects, with comparable performance to image-based approaches, by just utilizing WiFi signals. Analogous to Zhao et al. [Zha18b], the authors train a teacher-student network to address the missing annotated data for HPE in WiFi settings. Finally, the authors claim that using WiFi is a privacy-friendly, illumination-invariant, and cheap human sensor compared to RGB cameras.

Apart from using WiFi, HPE using Radar technology has garnered significant attention in recent research efforts. Three notable papers contribute to this field: Zhou et al. [Zho23] introduced a U-Net architecture called MD-Pose, a method designed for single-channel UWB Radar systems, aiming to estimate human poses accurately by leveraging the unique characteristics of UWB Radar signals, achieving promising results in HPE tasks. The models allows to extract velocity information of various human body parts from the

Micro-Doppler signature for up to 14 skeletal keypoints and reconstruct fine-grained human poses. Kim et al. [Kim23] presented a study on 3-dimensional HPE employing Impulse Radio Ultra-Wideband (IR-UWB) Radar in conjunction with a transformer-based deep learning model called 3D-TransPOSE. Lee et al. [Lee23] developed Human Pose with Millimeter Wave Radar (HuPR), a benchmark dataset tailored for 2-dimensional HPE utilizing Millimeter Wave Radar, serving as a valuable resource for researchers to evaluate and compare different HPE algorithms specifically designed for Radar-based systems. The authors investigate various methods to pre-process the obtained Radar-signal. Similar to the other work previously presented, annotation for the Radar-domain is obtained by using an off-the-shelf human pose estimator, namely HRNet [Sun19], pre-trained on MPII, a publicly available dataset. The pose data from the visual domain is then used to train a pose estimator in the Radar domain. They compare their results to [Zha18b], showing their supremacy with respect to the localization of keypoints.

These papers collectively contribute to the advancement of HPE through WiFi and Radar technology, introducing innovative methods, datasets, and benchmarks that facilitate more accurate and robust HPE in various real-world applications. All approaches based on WiFi signals and Radar have in common, that they were investigated just for single or few people at the same time, with all datasets and experimental setups being located in a lab-like environment. This makes it difficult to estimate how well such approaches would perform in-the-wild.

In summary, the field of HPE is broad and diverse, with different fields of application and hence various resulting methods. Depending on the field, certain approaches are more or less suited. Especially, in the field of CCTV classical hardware like RGB-cameras are predominantly used due to their availability. This makes in particular those approaches coming from the imaging sector, ranging within the electromagnetic spectrum from TIR, over NIR to the visible part of the spectrum most relevant to this work. With respect to the dimensionality of human poses, i.e., whether they are 2- or 3-dimensional, and the kind of human body model, there exists a clear bias towards 2-dimensional skeletons in the existing literature. This makes such especially interesting for

the challenging field of urban video surveillance applications, where low resolutions and small pedestrian sizes are just some examples for existing challenges. Therefore, for the rest of this thesis, the methods concentrate on the MPPE-scenario using RGB-sensors and 2-dimensional kinematic, i.e., skeletal, human body models. However, as stated earlier, the pipeline itself is not limited to such and depending on the application, other input-modalities can also be of interest.

3.1.2 Datasets

As in any other field, data is an essential resource for training models, developing methods and finally assessing their performance. Therefore, in the following different existing RGB datasets for HPE are presented, starting with the most prominent and widely used real-world datasets, followed by some synthetically generated ones.

3.1.2.1 Real-World Datasets

3.1.2.1.1 Common Objects in Context - Keypoints

Common Objects in Context (COCO) [Lin15] is a large scale dataset designed for tasks like object detection, segmentation, captioning, and keypoint detection. It consists of more than 330,000 images and contains segmentation and detection labels for 80 different types of objects. Beyond that, COCO also provides around a quarter million labeled human poses in very diverse settings focusing on single or few people per image. As illustrated in Figure 2.5, the poses consist of 17 keypoints $\mathbf{v} \in V$ with labeled 2-dimensional positions as well as a flag to describe each keypoints visibility as introduced in Section 2.2.1. The visibility flag is defined as

$$v^{(2)} := \begin{cases} 0 & \text{if } \mathbf{v} \text{ is not labeled} \\ 1 & \text{if } \mathbf{v} \text{ is labeled but not visible} \\ 2 & \text{if } \mathbf{v} \text{ is labeled and visible} \end{cases} \tag{3.1}$$

It is important to note that occluded keypoints that are within the image bounds can be flagged as either *not labeled* or *labeled but not visible*. This means that the keypoints that are flagged as *labeled but not visible* are only a subset of occluded keypoints. In defiance of the datasets popularity stands its biggest downside for the purpose of this work, namely its lack of keypoint annotations for crowded scenes. Specifically for the task of keypoint localization Ronchi et al. [Ron17] added a new metric to evaluate the precision and recall of the predicted keypoints called *Object Keypoint Similarity (OKS)*. The mentioned metric will be presented in Section 5.2.1.1 of this work.

3.1.2.1.2 MPII Human Pose Dataset

The MPII Human Pose Dataset (MPII) is a benchmark for the evaluation of articulated HPE introduced in [And14]. The dataset includes around 25,000 images containing over 40,000 people with annotated body joints. The images were systematically collected using an established taxonomy of every day human activities. Overall the dataset covers 410 human activities and each image is provided with an activity label. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. In addition, for the test set we obtained richer annotations including body part occlusions and 3-dimensional torso and head orientations. As mentioned in Section 2.2, there exists a variety of pose models. With the MPII dataset Andriluka et al. [And14] provide their own model consisting of 14 keypoints, i.e., three less than COCO. Though most of the keypoints of these two models refer to the same body parts such as legs and arms, there is still a difference between both. In particular, the COCO model represents a human's head with more detail compared to the model given in the MPII dataset. MPII has only one keypoint located on the head, which is a *head top* keypoint, however, COCO human pose model has five keypoints located on the head, such as ears, nose and eyes.

3.1.2.1.3 CrowdPose

The previously introduced datasets only provide keypoint annotations for non-crowded scenarios, which leads to the problem that current approaches do not account for dense human crowds. In order to alleviate this problem, Li et al. [Li19] recently published a public benchmark for pose estimation in crowded scenarios namely CrowdPose. The CrowdPose dataset, introduced in 2019, serves as a pivotal benchmark for pose estimation in crowded scenes. This dataset is specifically designed to address the challenges posed by crowded environments, where multiple individuals may occlude each other, making accurate pose estimation a complex task. The dataset provides a realistic and diverse set of images capturing various crowded scenarios, enabling researchers to develop and evaluate pose estimation algorithms under challenging conditions. The images in the dataset are collected in a way to achieve a uniform distribution of the density of human crowds. The aim was to prevent a bias of the dataset for crowded or sparse scenarios and to assure that approaches that perform well on this benchmark perform well on both crowded and non-crowded data. In addition, Li et al. [Li19] introduced the Crowd Index (CI), which is a measure for the crowdedness level within an image. It essentially maps a given situation to a score by taking each person and putting them into relation with other closely located pedestrians. A more detailed presentation of the CI is given later in Section 3.2.2. Li et al. [Li19] selected images for the dataset in such way, that a uniform distribution of the CI between $[0, 1]$ is achieved.

3.1.2.1.4 PoseTrack

The PoseTrack dataset is a significant contribution to the field of HPE and articulated tracking. It serves as a large-scale benchmark specifically designed for video-based HPE and tracking tasks, containing over 150,000 poses in more than 22,000 video frames [And18]. The videos are annotated densely

with body poses with 15 keypoints[1], a head bounding box and a tracking id for every person. This dataset addresses the need for comprehensive evaluation in complex event video scene analysis, providing researchers with a valuable resource to test and benchmark their algorithms in real-world scenarios. PoseTrack dataset includes challenging situations with highly occluded individuals and complex movements in crowded environments, making it a suitable testbed for assessing the robustness and accuracy of pose estimation and tracking models [And18]. While the CI distribution of PoseTrack shows that the dataset provides a better representation of human crowds than most other datasets, the underlying problem, i.e., the limited amount of video sequences, still remains. This limits the variety of scenarios and poses drastically that is why most contenders on the *PoseTrack Challenge* use additional data for training and only fine-tune on PoseTrack [Kal19].

3.1.2.1.5 Summary

Although there are several real-world datasets available, they all tackle rather similar scenarios, non of which represents the typical surveillance scenarios and their challenges. Furthermore, since all datasets are mainly recorded or collected in western countries, a certain bias for the western culture and appearance is present. This makes these datasets suited for pre-training, but lacks enough domain-specific information to solely rely on for training data-driven approaches for the field of video surveillance.

3.1.2.2 Synthetic Datasets: Joint Track Auto and Extension

Joint Track Auto (JTA) is a synthetic dataset for HPE and tracking which Fabbri et al. [Fab18] collected by extracting videos from the video game *Grand*

[1] Mostly identical to COCO, with a different representation of the head of a person. Can be seen as a mix between MPII and COCO [And18]

Theft Auto (GTA) V[1]. The advantage of synthetically acquired data is that annotating such is done easily by extracting relevant information directly from the game engine and therefore no manual labor is required for labeling. Different to most real-world datasets, JTA provides 22 annotated keypoints, alongside with tracking ids and visibility flags. Since previously mentioned datasets purely contain manually labeled image data, they do not provide reliable annotations for occluded keypoints. JTA on the other hand provides highly accurate annotations for all keypoints within an image by utilizing the game engine to extract information even of occluded keypoints. The provided visibility flags allow to distinguish between keypoints that are visible, occluded and self-occluded. This could enable pose estimation methods to predict even the position of occluded keypoints.

The scenes in JTA contain on average 21 pedestrians and up to 60 people. This is a significantly larger number compared to previously released datasets, where the usual number of people per frame ranges from 1 to 13 [Kal19]. While the majority of keypoint annotations are accurate, the given extraction method introduces certain systematic errors, which occur in combination with specific person models within the game. Since these errors occur rarely, no further action has been taken to remove them from the dataset. Firstly, the scenes include a much higher density of people per frame. Secondly, the crowding level of the scenes is also much higher than other datasets with the exception of *CrowdPose*. Third, the addition of occlusion flags can provide substantial advantage for the pose estimation for crowded scenarios. Finally, the tools for creating own sequences in GTA are also freely available, so it is possible to extend the dataset.

Some of the challenges mentioned above, like a non-uniform Crowd Index distribution or the limited pose variety, can be solved by utilizing the JTA-Mod Fabbri et al. [Fab18] developed for creating the JTA dataset. The JTA-Mod allows to create scenarios in different settings within the GTA game world.

[1] GTA V is an action-adventure video game developed by *Rocksstar North* which was first released in 2013. It comes with a large and strong modding community. More information on the video game can be found here: https://www.rockstargames.com/gta-v

In these settings pedestrians can be placed performing different activities and showing various behaviors. The original JTA dataset primarily focuses on walking pedestrians, which leads to low pose variety. This is why the extension provided by Golda et al. [Gol19b] includes all possible activities that can be generated with the JTA-Mod, such as sitting, doing yoga, push-ups, cheering and fighting.

Finally, in the original JTA dataset the number of poses for a Crowd Index above 0.5 was disproportionately low, so for the extension aims to create denser crowds and adding more people to the scene. The final extension includes 58 additional video sequences with a total of 46,350 frames and 812,742 poses. As stated in [Kal19, Gol19b], the addition of different pedestrian activities in the JTA-Extension helped to diversify the pose variety. While the deviation from the maxima is still low compared to CrowdPose, there is a noticeable improvement to the original JTA pose variety.

In conclusion, as Kalb [Kal19] showed, the Crowd Index distribution of JTA-Extension results in an improvement compared to the original JTA, but still lacks uniformity. A way to address this issue would be to sample the data so that it is uniformly distributed, just like Li et al. [Li19] did for the CrowdPose dataset. Nonetheless, since the distribution is close to being uniform, the images are not resampled. Additionally, sampling frames from video sequences the JTA-Mod provides could lead to some videos being over-represented in the training data.

3.2 Characterizing Crowds for Human Pose Estimation

As crowds or so called crowded situations are one major topic of this thesis, it is interesting to see how such can be characterized or measured in general. Therefore, this section focuses on crowd measures. Only few measures for characterizing crowds with a tight focus on pedestrians were proposed yet, namely rather static measure as Multi-Object Tracking (MOT) [Lea15]

and the CI [Li19], or such incorporating dynamics, like Crowd Collectiveness [Zho14] or Crowd Aggregation [Xu18]. The aim of the first two is to describe the difficulty of a given scene based on information about the number and distribution of pedestrians. None of the existing measures is able to deliver an intuitive way to describe how difficult a given sequence (on frame level) actually is. Let $\mathcal{I} \subset \mathbb{N}_0$ be the set of pedestrians in the scene with $n = |\mathcal{I}|$ for all following measures.

3.2.1 Multi-Object Tracking

The first measure is one introduced at the MOT challenge and is built very naively. The density of a scene is defined as the average number of people per frame for a given sequence.

$$f_{\text{MOT}} := \frac{1}{n_{\mathcal{F}}} \sum_{k \in \mathcal{F}} \nu_k, \tag{3.2}$$

where $k \in \mathcal{F}$ is an index for a single frame from sequence $\mathcal{F}$, $n_{\mathcal{F}} = |\mathcal{F}|$ is the total number of frames in the sequence and $\nu_k \in \mathbb{N}_0$ is the number of pedestrians in frame k. However, this lacks any information about the distribution of the people, i.e., whether they are closely gathered or spread across the frame [Lea15].

3.2.2 Crowd Index

Different to the MOT measure, the Crowd Index [Li19] takes the spatial relation of each person to surrounding persons into account. It was initially developed for HPE purpose and requires pose information for all pedestrians within the sequence in order to get computed. The CI for a given image is obtained by following formula

$$f_{\text{CI}} := \frac{1}{n} \sum_{i=0}^{n-1} \xi_i, \tag{3.3}$$

where ξ_i is the Crowd Ratio (CR) for the ith person. In its original version the Crowd Ratio ξ_i for person i describes the ratio between two distinct sets of keypoints within the area of person's bounding box: First, the number of keypoints n_i^i within the bounding box of person i actually belonging to person i and second, the total number $\sum_{k \neq i} n_i^k$ within the same bounding box of those not belonging to person i.

$$\xi_i := \frac{\sum_{k \neq i} n_i^k}{n_i^i} \tag{3.4}$$

3.2.3 Crowd Collectiveness and Aggregation

Crowd Collectiveness and Crowd Aggregation are fundamental concepts in the study of crowds and their behaviors. Crowd Collectiveness, as defined by Zhou et al. [Zho14], refers to a universal metric that quantitatively measures the level of cohesion or dispersion within a crowd. It provides insights into the collective organization of individuals within a crowd, shedding light on how groups of people interact and move together. This is described as given in Equation (3.5)

$$\Phi := \frac{1}{n}\mathbf{e}^\intercal\big((\mathbf{I} - z \cdot \mathbf{W})^{-1} - \mathbf{I}\big)\mathbf{e} \tag{3.5}$$

where, $\mathbf{e} := (1, \ldots, 1) \in \mathbb{R}^n$ and $\mathbf{W} \in \mathbb{R}^{n \times n}$ being the weighted adjacency matrix of the graph, where an edge $w^{(i,j)}$ is the similarity between individual i and j in its neighborhood. Furthermore $z < 1$ is a real-valued regularization factor [Zho14]. On the other hand, Crowd Aggregation, as introduced by Xu et al. [Xu18], focuses on the efficient computation of crowd density and distribution in public areas. It aims to analyze how individuals gather and disperse in crowded spaces, providing valuable information for various applications such as crowd management and urban planning.

The concepts of Crowd Collectiveness and Crowd Aggregation are closely related in that they both deal with the spatial and social dynamics of crowds. In summary, Crowd Collectiveness and Crowd Aggregation are interconnected

concepts that provide valuable insights into the behavior and organization of crowds. They are computed through analyzing spatial dynamics, movement patterns, and density distribution within crowds, offering essential information for various fields such as crowd simulation, urban planning, and social behavior analysis. With respect to the task of HPE, they are less useful and although they aim to describe a crowd, both measures are less suited to assess how challenging a given situation is.

3.3 Domain Adaptation

In the realm of domain adaptation various surveys [Csu17a, Zha18c, Wan18, Wil20, Oza21] offer an overview of methods for addressing domain adaptation for computer vision tasks each of which addresses the field from slightly different perspectives. For instance, Oza et al. [Oza21] focus on the challenges and techniques related to adapting object detectors to new domains without labeled target data. Whereas Csurka [Csu17a] provides an additional glimpse on transfer learning. Together with the work of Zhao et al. [Zha18c], Wang et al. [Wan18], and Wilson et al. [Wil20] the surveys provide a comprehensive view of domain adaptation, covering various aspects such as object detection, visual applications, deep learning methods, and unsupervised adaptation techniques.

This part sheds light on some of the commonly chosen techniques and methods that are used within this thesis, as Domain Adaption (DA) is used to alter synthetically generated data in order to generate even more suited training data in an privacy-friendly way.

3.3.1 Cycle-GAN

The Cycle-GAN [Zhu17] approach introduces two distinct generator networks $G_{st}(\mathbf{x}_s) : \mathcal{D}^s \rightarrow \mathcal{D}^t$ and $G_{ts}(\mathbf{x}_t) : \mathcal{D}^t \rightarrow \mathcal{D}^s$, that map from either source to target domain or target to source domain. In addition, for each of the two domains there is a discriminator network $D_t(\mathbf{x})$ and $D_s(\mathbf{x})$, that

categorize between real and fake samples. Thus, there are two GAN frameworks, one for each domain that are both trained using the classical GAN objective. The resulting objective for the domain adaptation from source to target domain, and from target to source domain are given in Equation (3.6) and Equation (3.7) respectively.

$$\begin{aligned}\min_{G_{st}} \max_{D_t} \mathcal{L}_{G,s\to t}(G_{st}, D_t) = \mathbb{E}_{\mathbf{x}_t \sim p(\boldsymbol{X}_t)}[\log D_t(\mathbf{x}_t)] \\ + \mathbb{E}_{\mathbf{x}_s \sim p(\boldsymbol{X}_s)}[1 - \log D_t(G_{st}(\mathbf{x}_s))]\,,\end{aligned} \tag{3.6}$$

$$\begin{aligned}\min_{G_{ts}} \max_{D_s} \mathcal{L}_{G,t\to s}(G_{ts}, D_s) = \mathbb{E}_{\mathbf{x}_s \sim p(\boldsymbol{X}_s)}[\log D_s(\mathbf{x}_s)] \\ + \mathbb{E}_{\mathbf{x}_t \sim p(\boldsymbol{X}_t)}[1 - \log D_s(G_{ts}(\mathbf{x}_t))]\,,\end{aligned} \tag{3.7}$$

Note that for the original version, the objectives are formulated as mean squared error losses and therefore differ from the Equation (3.6) and Equation (3.7).But since the objective functions depicted above, together with the one sided gradient penalty as presented in Section 2.1.5, have been found to provide more stability for the training runs, that were conducted within this thesis, they are applied to all of the utilized Cycle-GAN models. For more information on the original Cycle-GAN formulation see [Zhu17].

As already stated, it is intended to reduce the space of possible mappings $G_{st}(\mathbf{x}_s)$ and $G_{ts}(\mathbf{x}_t)$, and thus to facilitate the preservation of the content for adapted input images. Such a reduction in dimensionality can be achieved by transforming adapted samples back to the domains they originate from and requiring the output of this retransformation to be identical to the original sample, which is expressed by the following two formulas:

$$G_{ts}(G_{st}(\mathbf{x}_s)) \overset{!}{=} \mathbf{x}_s \tag{3.8}$$

$$G_{st}(G_{ts}(\mathbf{x}_t)) \overset{!}{=} \mathbf{x}_t \tag{3.9}$$

3.3.2 Cycle-Consistent Adversarial Domain Adaptation

Cycle-Consistent Adversarial Domain Adaptation (CYCADA) [Hof17] extends the Cycle-GAN especially for domain adaptation by proposing to incorporate a task loss to ensure semantic consistency before and after domain adaptation. This achieved promising results from both qualitative and quantitative perspective.

The first part of the CYCADA method is to train a discriminative task model $f_{\mathcal{T}}$ in the source domain, where labels are available. Naturally, the specific form of the task loss depends on the task itself. In the case of a C-way classification problem this task loss could be

$$\min_{f_{\mathcal{T}}} \mathcal{L}_{\text{task}}(f_{\mathcal{T}}, \boldsymbol{X}_S, \boldsymbol{Y}_S) = \mathbb{E}_{\{\mathbf{x}_S, y_S\} \sim p(\boldsymbol{X}_S, \boldsymbol{Y}_S)} \left[\sum_{c=0}^{C-1} \ell_c(\mathbf{x}_S, y_S) \right], \tag{3.10}$$

with

$$\ell_c(\mathbf{x}_S, y_S) \coloneqq \mathbf{1}_c(y_S) \cdot \log\left(\sigma(f_{\mathcal{T}}(\mathbf{x}_S))\right) \tag{3.11}$$

where y_S is the ground truth label to an image $\mathbf{x}_S$ from the source domain, $\mathbf{1}_c : Y \to \{0, 1\}$ is the indicator function with

$$\mathbf{1}_c(y) \coloneqq \begin{cases} 1 & \text{if } y = c \\ 0 & \text{if } y \neq c \end{cases} \tag{3.12}$$

and $\sigma(\mathbf{x})$ denotes the softmax function. Generally speaking, $\mathcal{L}_{\text{task}}$ can be any kind of objective. However, for all models within this thesis, the utilized loss function is exactly the one depicted in Equation (3.10). This loss should point the generator the way to producing images that are semantically meaningful.

During the training of the Cycle-GAN, the learned model $f_{\mathcal{T}}$ is applied to samples from both domains $\mathbf{x}_S$ and $\mathbf{x}_t$, and the resulting adaptations $G_{st}(\mathbf{x}_S)$ and $G_{ts}(\mathbf{x}_t)$ respectively. This way, both generators are enforced to produce images yielding the same classification result throughout this process, hence ensuring that both, original input and the adapted output, are semantically

consistent. The resulting loss is as follows:

$$\begin{aligned}\min_{G_{st},G_{ts}} \mathcal{L}_{\text{sem}}(G_{st}, G_{ts}, f_{\mathcal{T}}, \boldsymbol{X}_s, \boldsymbol{X}_t, \boldsymbol{Y}_s) = \mathcal{L}_{\text{task}}\Big(f_{\mathcal{T}}, G_{st}(\boldsymbol{X}_s), \boldsymbol{Y}_s\Big) \\ + \mathcal{L}_{\text{task}}\Big(f_{\mathcal{T}}, G_{ts}(\boldsymbol{X}_t), \hat{\boldsymbol{Y}}_t\Big),\end{aligned} \tag{3.13}$$

where $\hat{\boldsymbol{Y}}_t \coloneqq \arg\max(f_{\mathcal{T}}(\boldsymbol{X}_t))$, i.e., the labels for the second part of Equation (3.13) are the predictions $f_{\mathcal{T}}(\boldsymbol{X}_t)$ that are most likely. This is made because no target labels are available. The loss is used as training objective in addition to the GAN losses from Equations (3.6) and (3.7), and the *cycle-consistency loss* from Equation (4.16).

Note that the additional losses that are introduced in Equations (3.10) and (3.13) are only the parts of the CYCADA method that affect pixel space adaptation, i.e., the adaptation of training images. The subsequent parts are here skipped as they consider the classifier that is to be learned from the adapted data. For more detailed information on CYCADA see [Hof17].

3.4 Behavioral Anomaly Detection

The task of recognizing anomalous behavior in the field of public safety attracted researchers for many years. As stated by Amrish et al. [Amr23] there are various ways to assess the level of anomalousness of a given scene from a human-centered perspective. These are: structural information, like the number of people and their distribution, and behavioral information, encompassing overall motion, actions and activities. From this human-centered perspective the behavioral anomalies are of special interest to this thesis, which stands in contrast to classical action or activity recognition tasks. Therefore, the following pages will give an overview over methods and ways to deal with such.

The field of anomaly detection is dominated by two kinds of approaches, namely *reconstruction-* and *prediction-based* [Li20]. Reconstruction-based methods aim to reconstruct input video frames with minimal errors for

normal situations and detect anomalies based on significantly higher reconstruction errors for abnormal situations. These approaches rely on the assumption that normal examples can be accurately reconstructed from a latent representation. They are effective in representing normal patterns but can be limited by heavy reliance on training data [Ris21, Bia21].

On the other hand, prediction-based approaches focus on predicting future steps and identifying anomalies by comparing the predicted with the actual data. These methods are sensitive to noise in complex scenarios and can detect anomalies spatially and temporally by analyzing prediction errors of normal and abnormal trajectories [Kan22].

Despite their flaws, both ways are commonly applied to the different fields, especially those presented in the following.

3.4.1 Video-based Anomaly Detection

Video-based Anomaly Detection (VAD) is an area within computer science that focuses on identifying abnormal events, patterns or behaviors in video data. It involves algorithms and models to differentiate between normal and anomalous activities captured in video streams. Anomaly detection in videos is a challenging task due to the diverse nature of anomalies present in different videos. The goal is to automatically detect deviations from expected behaviors, which can range from simple anomalies to complex events that require sophisticated analysis [Li20]. Many of these approaches do not solely concentrate on humans, but rather take any kind of anomalous events or patterns into account, which makes them much more versatile. By leveraging advanced algorithms, deep learning models, and spatio-temporal features, existing methods aim to be capable of detecting a wide range of anomalies in various domains. Such might include traffic, certain objects or, closer human-related crowds and their distribution or dynamic [Ald22]. In the context of behavioral anomalies, VAD can be simply summarized as methods that take the overall motion into account, either directly [Rav17a, Rav17b, Rei22] or indirectly [Wan22].

One particular type of methods that follow such a *macroscopic* approach are for example GAN-based solutions, like those by Ravanbakhsh et al. [Rav17a, Rav17b], Lee et al. [Lee18], and Singh et al. [Sin23]. Due to their generative character, they are typically used in predictive but also reconstructive manner. Ravanbakhsh et al. [Rav17a, Rav17b] proposed to use GANs, which are trained using normal frames and corresponding optical flow (OF) images in order to learn an internal representation of the scene normality. Since they are trained with normal data only, they are not able to generate abnormal events. During inference, both the appearance and the motion representations reconstructed by these GANs are compared to the real data and abnormal areas are detected by computing local differences.

Lee et al. [Lee18] on the other hand proposed a spatio-temporal generator that generates inter-frames by considering neighboring frames. Then, consecutive frames including the generated inter-frame are considered as a fake sequence while a real sequence contains only real frames. Through the adversarial training, the spatio-temporal discriminator learns to determine whether the given input sequence is real or fake, while the generator is trained to generate an inter-frame that can fool the discriminator.

Another very similar approach was proposed by Singh et al. [Sin23]. They introduced their model called Spatio-Temporal Generative Adversarial Network (STemGAN), which consists of a generator and discriminator that learn from the video context, utilizing both spatial and temporal information to predict future frames. The generator is an AE-architecture, with a dual-stream encoder for extracting appearance and motion information, and a decoder having a Channel Attention (CA) module to focus on dynamic foreground features. In addition, they provide a transfer-learning method that enhances the ability of their STemGAN to generalize.

In conclusion, the presented GANs-based approaches follow mostly similar ideas with few significant differences. GANs are known to be difficult to train, which is one reason that many alternative types of approaches prevail.

Apart from the GAN-based approaches, various other holistic methods exist. Georgescu et al. [Geo21] for instance presented an object-oriented framework

for abnormal event detection in videos, equipping a background-agnostic approach with adversarial training. They present a three-fold AE-architecture that aims on reconstructing given inputs. It consists out of two types of AEs one for appearance of the frame at timestep t and the other one for motion, represented as the OF between timesteps $t-1$ and t, as well as t and $t+1$. During training, each of the AEs gets adversarial pseudo-abnormal samples to reconstruct, following the idea, that the binary classifier generates a lower score if its abnormal and a higher score if its normal. The appearance AE is additionally trained with segmentation masks in order to focus on objects and pedestrians. The authors state, that the adversarial part brings major improvement to their approach.

Wang et al. [Wan22] proposed a method for VAD by solving decoupled spatio-temporal jigsaw puzzles. The essence of their idea is to learn a model, which is capable of finding a permutation that de-shuffles the previously shuffled input volume. The mentioned volume consists of a stack of frames, which is split into 3-dimensional blocks that are shuffled in two ways: temporally and spatially, resulting in a temporal and a spatial permutation. Finally, the network gets both "puzzles" and has to generate a permutation matrix that reconstructs the initial input. The permutation matrix is then used to generate a score for anomaly ranking.

Following a simple idea, Reiss et al. [Rei22] focused on attribute-based representations for accurate and interpretable VAD. The authors extract three kinds of features: velocity of the movement within the scene based on OF, human poses and deep features using Contrastive Language-Image Pretraining (CLIP). The last two are extracted from bounding boxes that are generated in the first place using an off-the-shelf object detector like YOLO. Finally, for each feature a density estimate is computed, which is used to classify incoming samples as anomalous, if the density-score is low, and normal otherwise.

The final approach by Liu et al. [Liu23] propose a Diversity-Measurable Anomaly Detection (DMAD), which is a reconstruction-based framework to enhance the measurability of reconstruction diversity so as to measure abnormality more accurately. The basic idea is to decouple the reconstruction into compact representation of prototypical normals and measurable deformations

of more diverse normals and anomalies. The under-estimated reconstruction error can be compensated by the diversity, which can be properly measured. To this end, the DMAD framework includes a pyramid deformation module to model and measure the diversity and an information compression module to learn the prototypical normal patterns. The authors assume anomalies can be represented as significant deformation of appearances, including positional changes and fine motions. In contrast, diverse normal samples can be represented as weaker deformations thus easily distinguished.

All these methods aim on solving the general idea of anomaly detection in videos, where humans are just a singular possible reason for the occurrence of anomalies. Therefore, data-privacy is a scarcely addressed topic and not regarded in any of the above mentioned examples. The next section eliminates this holistic strategy and presents so-called *microscopic* methods that directly target human beings.

3.4.2 Skeleton-based Anomaly Detection

For this section on related work in the field of Skeleton-based Anomaly Detection (SBAD) a brief introduction to existing approaches is provided. As stated in Section 2.2, the phrase *skeleton* corresponds to the 2-dimensional human pose models, which serve as exclusive input to the behavior analysis. Since the aspect of privacy-friendliness plays an important role, this section focuses solely on methods that are just based on the structural and temporal information contained in skeleton or pose sequences. Established and widely utilized in anomaly detection for vision-based applications, unsupervised learning approaches have proven their effectiveness. Such methods exclusively train on normal videos and excel at identifying anomalies that deviate significantly from typical behavior. Ongoing research in this domain has yielded various techniques, demonstrating high accuracy and effectiveness across diverse environments.

The Message-Passing Encoder-Decoder Recurrent Neural Network (MPED-RNN) introduced by Morais et al. [Mor19] comprises two separate Recurrent Neural Network (RNN) branches, each dedicated to global and local feature

components. These branches process data independently and communicate through cross-branch message-passing at each time step. The model is trained end-to-end with regularization, aiming to distill a concise profile of normal training patterns for efficient detection of abnormal events in a unsupervised learning manner.

Another innovative method, the Graph Embedded Pose Clustering (GEPC) by Markovitz et al. [Mar20], detects anomalous behaviors by utilizing an AE and clustering, samples are mapped to a latent space with soft clustering, leading to the creation of a bag-of-words representation for actions. The soft-assignment vector distribution is measured by the Dirichlet Process Mixture Model (DPMM), enabling the determination of normality scores for action classification.

Furthermore, the work of Rodrigues et al. [Rod20] introduces a multi-timescale framework comprising two models: one predicting future sequences and the other inferring past sequences based on single-person pose sequences. Supervision at intermediate layers enables the model to learn temporal dynamics at multiple timescales, addressing the problem of abnormal human activity detection effectively.

The baseline model called Normal Graph by Luo et al. [Luo21] employs a prediction approach based on Spatial-Temporal Graph Convolutional Networks (ST-GCNs) presented in Yan et al. [Yan18]. Multiple ST-GCN layers progressively deliver joint information, concluding with a prediction module for forecasting future skeleton configurations. Analyzing prediction errors enables the identification of normal or abnormal behavior based on normalized scores.

Furthermore, Liu et al. [Liu21a] introduce the Spatial Temporal Self-attention Augmented Graph Convolutional Autoencoder (SAA-STGCAE). This method first extracts poses and employs the AE of SAA-STGCAE, followed by the DPMM, adopted from [Mar20], to generate a normality score and identify outliers.

The Hierarchical Spatio-Temporal Graph Convolutional Neural Network (HSTGCNN) proposed by Zeng et al. [Zen23] comprises three main components: a spatio-temporal graphical feature extractor, a future trajectory

predictor and an outlier arbiter. It processes inputs by organizing them into a spatio-temporal graph representing human body joints across multiple frames. The model is trained on normal activities, allowing it to accurately predict trajectories for human joints during routine behaviors. Different branches of the model handle various scenes, leveraging weighted combinations of graph representations. Optical flow fields and average sizes of human bounding boxes and skeletons are used to cluster videos into scene-specific groups, determining branch weights. The outlier arbiter then combines predictions from different branches to generate the final anomaly score.

The Spatio-Temporal Graph Normalizing Flows (STG-NF) model proposed by Hirschorn et al. [Hir23] comes with a lightweight architecture (about 1,000 parameters) that employs normalizing flows in a spatio-temporal pose data framework. In the first step of the framework, a sequence of video frames is taken as input. Next, human poses are extracted of each person in every frame and a pose tracker is used to trace the skeletons over time. Ultimately, each pedestrian is represented as a temporal pose graph per video clip. The training samples are mapped into a Gaussian-distributed latent space by the STG-NF network and the probability of a human pose sequence is calculated. Lastly, Motion Prior Regularity Learner (MoPRL), proposed by Yu et al. [Yu24], addresses the lack of direct dynamic representation by introducing the Motion Embedder (ME). Serving as a label-efficient scheme, ME offers a probabilistic perspective on pose motion for structured data, eliminating the need for extra annotations. Additionally, Yu et al. [Yu24] introduced the task-specific Spatial-Temporal Transformer (STT) to enhance the model. Designed for self-supervised pose sequence reconstruction, the integration of ME and STT in a unified framework forms the powerful MoPRL, providing a comprehensive solution for pose regularity learning.

4 Methodical Considerations

4.1 Overview

As the aim of this thesis is to investigate how privacy-friendliness can be implemented with respect to human behavior analysis, and in particular recognition of salient or anomalous behavior, this chapter presents different approaches that are introduced in detail. The preceding chapters already indicated that a central feature chosen for implementing such behavior are skeletons representing human poses, which already accomplish to reduce privacy concerns in modern smart surveillance system as shown in a recent study [Gol22a]. However, as argued in Section 1.6 the overall pipeline around an approach founded on skeletons is also prone to possibly privacy-related issues. This chapter therefore presents various attempts to address several of these concerns. Section 4.2 begins with a look on synthetic data as an alternative for real-world data, and addresses also the topic of getting a sufficient amount of data, since the acquisition of data for training models for HPE is difficult and time-consuming. Working with synthetic data generates a certain bias and moreover a domain gap that is crucial for the performance of the trained model. This issue is tackled in Section 4.3, which extends the work by addressing the mentioned domain gap between synthetic and real data using techniques known as domain adaptation. Finally, Section 4.4 focuses on the methodical part of analyzing video-based behavior analysis in two different ways, namely a macroscopic approach based on a GAN-based architecture and a microscopic based on the earlier mentioned skeletal human representation.

4.2 Synthetic Datasets

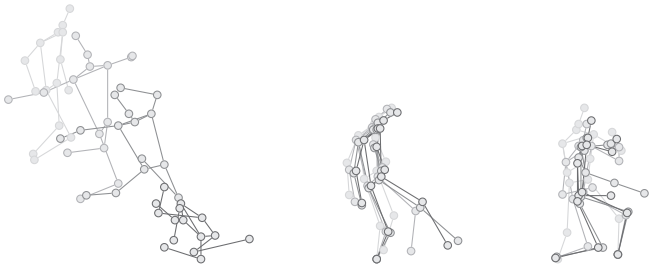
One way to address privacy concerns during the early stage of the development process is by avoiding to use sensitive data for training data-driven models. This can be achieved by considering just artificially generated data for training neural networks. With respect to the overall behavior analysis pipeline as introduced in Section 1.6, synthetic data can be beneficial for multiple tasks, namely training of the behavior analysis module, as well as a human pose estimator. The following sections show possible ways to generate training data for both, behavior recognition (Section 4.2.1), and human pose estimation (Section 4.2.2).

4.2.1 MixAMoR

Databases like Mixamo[1] by Adobe offer animated samples for animators and game developers, offering various kinds of behaviors and actions. In particular, Mixamo's database consists of motion-captured and post-processed animations. However, using such data is a rather time-consuming and complex approach to generate training samples that comes with several restrictions. The most prominent ones are the small number of possible variations in movement due to the manual recording and the short sequences, the difficult and inconsistent parametrization of available animations and the dominance of exaggerated animated actions, which are often far from reality. Last but not least, extracting data from the database and preprocessing it involves a lot of manual work, which is very time consuming and prone to errors. Katovich [Kat19] provides an extensive examination of Adobe's Mixamo database for the purpose of human behavior recognition, presenting not only information about the database but offering a Python-based plug-in for the open-source 3D software Blender[2]. Using this plug-in, this thesis at hand introduces the Mixamo Anomalous Movement Recognition (MixAMoR) dataset, a custom

[1] https://www.mixamo.com
[2] https://www.blender.org

tailored synthetic single-person salient behavior dataset. Based on the existing types of actions a suitable subset was collected by hand consisting of five activities, which are: *kicking*, *punching*, *running*, *tripping*, and *walking*. Other available activities were ignored mostly due to their unnatural appearance and involvement of weapons like swords or polearms. The resulting dataset consists of several short pose sequences[1], with an average length of 2.5 ± 1.3 seconds and ranging between 0.64 - 7.08 seconds that were exported from five different perspectives in 2-dimensional image coordinates. Figure 4.1 shows examples for three of these classes. No rendered frames are included in the dataset, as the data provided by Mixamo concentrates on the animation itself. As shown in Table 4.1, this results in a total of 750 sequences divided evenly into the five mentioned classes. Note that the MixAMoR dataset is used as it is and the training samples are directly extracted from the included behavior samples. However, the actions and movements collected from the Mixamo database could be potentially concatenated to generate longer sequences. This step comes with certain drawbacks though, especially with respect to the overall perception of the resulting sequences. Typically these sequences show varying inconsistencies with respect to the overall motion flow, since there is no guarantee that the involved movements can be connected seamlessly.



(a) Tripping **(b)** Kick **(c)** Punch

Figure 4.1: Examples for three different classes represented in MixAMoR. For better visualization every 6th timestep is drawn, which also shows how diverse the sequences are with respect to the duration and spatial extent. Each timestep is shown in a different shade of gray with darker shades indicating more recent states.

[1] The extracted raw data is available in FBX format, which stores the original 3-dimensional information and allows to export data from even more views.

Table 4.1: Characteristics for the MixAMoR dataset. The dataset consists out of 30 different sequences for each of the five classes. Each of these sequences was exported as 2-dimensional coordinates from five different views, resulting in 150 sequences per class.

	anomalous				normal
	kick	**punch**	**run**	**trip**	**walk**
train	132	107	121	117	123
test	18	43	29	33	27
total	150	150	150	150	150

Starke et al. [Sta22] showed that some of these drawbacks can potentially be diminished, by applying further AI-based approaches to increase the quality of the perceived motion for such concatenated animations. Especially when merging multiple movements into one consistent flow the approach can generate many more suitable training samples by fading one action into another with respect to the overall dynamics of the combined actions.

4.2.2 SyMPose

Different to the generation of training data for skeleton-based behavior recognition, training a human pose estimator requires labeled image data. Although, doing so with data provided by Mixamo would be possible as well, by animating various avatars using the obtained skeletal body models, the overall amount of work to prepare the data would require an exorbitant amount of time to generate a suited dataset.

As presented earlier in Section 3.1.2.2, Fabbri et al. [Fab18] introduced a dataset named JTA along with the source code for creating own datasets, which was intentionally designed as a Human Pose Estimation dataset. The authors took advantage of the large and active modding community of GTA, which made it easy to extract useful information for the task of training data generation. This procedure has already been applied to different problems like object detection [Kie21] and crowd counting [Wan19]. To overcome the limitations of JTA with regard to video surveillance applications for crowded scenarios, the Synthetic Semantic Maps and Human Poses (SyMPose) dataset has been collected [Bla19, Gol20] and used for various experiments throughout

this thesis. SyMPose is a dataset that focuses on video surveillance scenarios for crowded situations and consists of 21 video sequences[1] recorded in 1080p containing an overall number of 18,900 frames with 1,277,814 annotated keypoints. The number of people per frame varies between 22 and 130 and takes an average value of 68 pedestrians per frame. Due to the way SyMPose is designed, the overall amount of mutual and self-occlusions is as well much higher than those in JTA. Motivated by the surveillance setup, all sequences were recorded from differing overhead camera positions. Some examples from SyMPose are depicted in Figure 4.2. For further examples please refer to the work of Blattmann [Bla19]. As stated in his work, all sequences were collected in such way that their CI distributions differ. The same holds true for the Standard Graph Crowd Index (sGCI), due to the properties and the design of the sGCI. Further information on the sGCI is provided in Section 4.2.3.

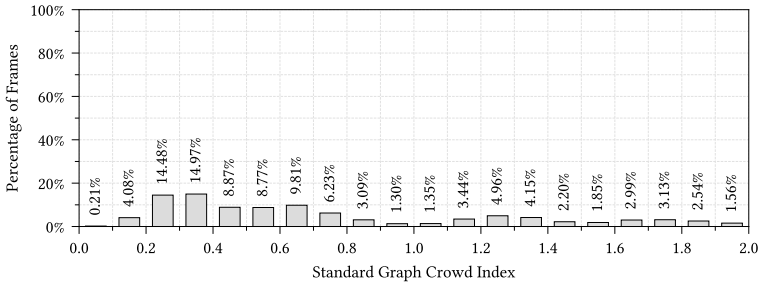


Figure 4.3: The overall distribution of the sGCI shows, that the crowdedness is approximately uniformly distributed over all frames of the dataset. An ideal balancing can be obtained by manually sampling a subset to generate an even more balanced statistic. However, the desired property of having temporal information and sequences makes it hard to reach the ideal case.

Figure 4.3 shows the resulting overall sGCI distribution. It is easy to see, that the sGCI is not ideally uniformly distributed, which is mainly accounted for by the temporal order of the single frames. One approach to operationalize

[1] Originally 25 sequences that were numbered from 1 to 25, four of them had to be removed due to artifacts.

(a) Sequence 5

(b) Sequence 8

(c) Sequence 12

(d) Sequence 15

(e) Sequence 17

(f) Sequence 25

Figure 4.2: SyMPose offers various settings, defined by their location, daytime and weather conditions as well as the number of pedestrians. All sequences have in common, that they were recorded in urban scenarios including inner-city areas, parks, or similar places. GTA offers many more possible places like deserts and other rural settings. These, however, would go against the the desired purpose of SyMPose.

this statement is by computing an arbitrary distance measure. When comparing CI and sGCI using the first Wasserstein distance W_1, it becomes evident that the latter is marginally closer to a uniform distribution than the former. This is highlighted by the values $W_1(P_{\text{CI}}, P_{\text{uniform}}) \approx 0.0499$ and

Table 4.2: Statistics for each video sequences included in the SyMPose dataset. μ_{sGCI} is the average value of the Standard Graph Crowd Index, μ_{ppf} and σ_{ppf} denote the average number of people per frame and the corresponding standard deviation respectively. Finally, the range of pedestrian counts per frame is indicated by r_{ppf}. The sGCI distributions for the listed sequences are plotted in Figure 4.5.

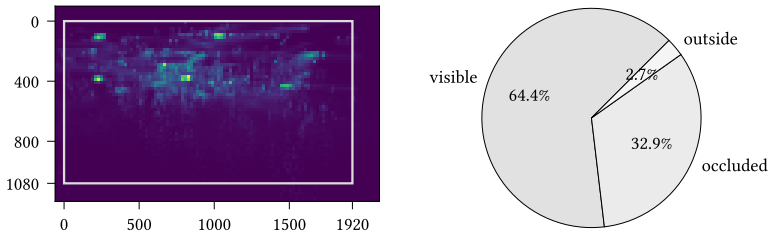
	μ_{sGCI}	μ_{ppf}	σ_{ppf}	r_{ppf}		μ_{sGCI}	μ_{ppf}	σ_{ppf}	r_{ppf}
S_1	1.23	73	2.30	68 - 78	S_{12}	0.69	56	3.93	43 - 62
S_2	1.35	81	1.44	79 - 85	S_{13}	0.26	38	2.02	34 - 41
S_3	0.47	53	5.93	40 - 64	S_{14}	0.28	64	4.28	57 - 72
S_4	0.42	55	9.69	45 - 79	S_{15}	0.71	77	4.92	67 - 86
S_5	0.33	46	2.77	39 - 53	S_{16}	0.37	57	3.71	52 - 65
S_6	0.58	62	9.40	40 - 74	S_{17}	0.32	54	3.25	46 - 61
S_7	1.35	75	7.17	62 - 88	S_{18}	1.88	95	2.95	89 - 102
S_8	1.72	115	6.24	106 - 130	S_{23}	0.64	45	7.48	30 - 61
S_9	0.21	27	3.67	22 - 36	S_{24}	0.74	88	6.09	76 - 97
S_{10}	0.49	63	5.19	55 - 73	S_{25}	1.50	91	9.53	74 - 107
S_{11}	0.53	92	2.35	87 - 96					

$W_1(P_{\text{sGCI}}, P_{\text{uniform}}) \approx 0.0331$ respectively. Here, P_i represents the discrete probability distribution resulting from the histograms. However, creating a balanced set out of the available frames can be easily achieved based on the source code provided by Blattmann [Bla19]. Table 4.2 and Figure 4.4 provide additional insight in the dataset's content with respect to the location and visibility of the keypoint annotations. The focus of SyMPose on crowded situations is evident given the statistics displayed in Table 4.2, and Figures 4.3 to 4.5. This makes the dataset suitable for research on and development of crowd-oriented approaches, especially for answering the research questions tackled in this thesis. For further information on SyMPose including additional statistics and examples please refer to Blattmann [Bla19] and Appendix A.1.

4.2.3 Graph Crowd Index

The in Section 3.2.2 introduced Crowd Index was developed for evaluation on ground level, which potentially comes with a larger difference in occupied image area between people close to the camera and those further away. Such cases have a strong influence on the CI. However, in the field of video surveillance and crowd monitoring the promised effects do not hold, since typical setups use cameras that are mounted relatively high above ground

level in overhead position. This results in a more even distribution of pedestrian sizes, which in general yields lower CI scores for such perspectives and hence does not adequately represent the difficulty of the scenario. Furthermore, the CI requires a dataset to provide pose information for every human represented within the dataset. This is a very restrictive pre-condition, since pose information is often expensive to collect [Cor21] and depending on the use case a complete annotation of all pedestrians with pose information is not available. As stated in Section 3.2, there exist only few metrics that aim on characterizing crowds and crowded situations of which the CI appears to be the fit as it was developed right for this kind of data, namely MPPE. This thesis presents a generalized version of the CI in order to address the shortcomings of the initial metric and especially to enable the description and characterization of arbitrary datasets with the provided metric. The essential idea for the newly presented family of metrics called Graph Crowd Index (GCI) is a person-centered graph representation of a given scene, which generalizes the CI by omitting the restriction to keypoint coordinates and just using the areas of interfering bounding boxes. In the following, the general concept of the GCI is introduced and extended.



(a) Spatial Histogram of Keypoint Locations

(b) Proportions of Annotated Keypoints on Occlusion-level

Figure 4.4: The histogram on the left visualizes the spatial distribution of keypoints all over every frame from SyMPose. Within the white rectangle, i.e., the actual visible part of the camera view, 97.3% of all visible and occluded keypoints are located. The remaining 2.7% are located outside the visible area. These categories are displayed on the right.

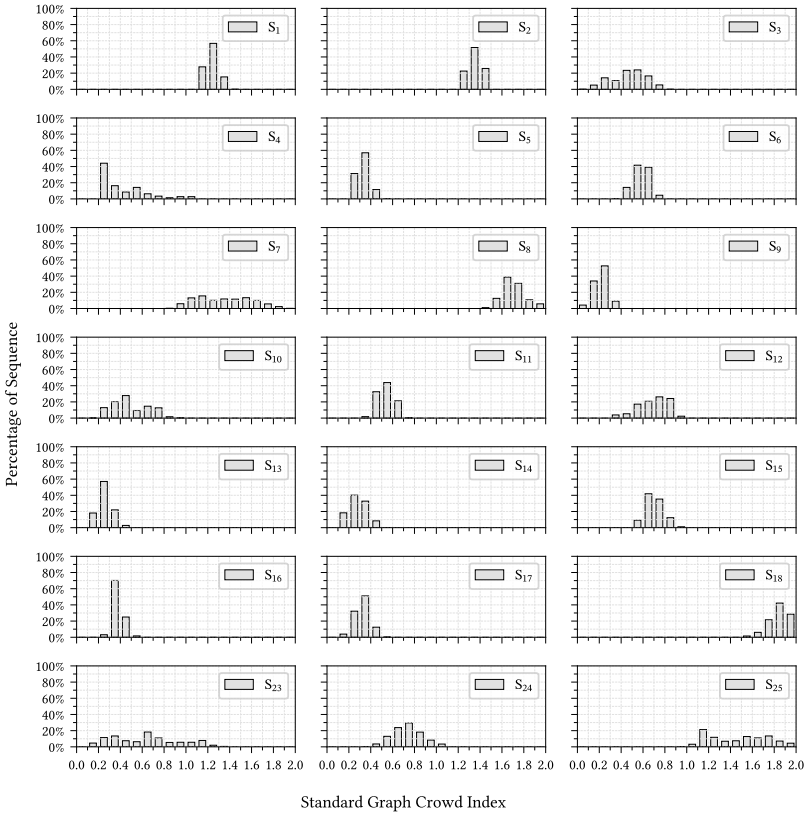


Figure 4.5: Overview over all index sGCI distributions for each sequence contained in the final SyMPose dataset. The Standard Graph Crowd Index eliminates the shift towards higher index values and hence, leads to a similar distribution as the original CI metric.

4.2.3.1 Definition

As mentioned above the newly presented metric[1] is based on a graph representation of the given scene. Therefore, a weighted connection graph between all

[1] Note that both the original Crowd Index and the proposed Graph Crowd Index are not metrics in the manner of the typical axioms which are defined for metric spaces.

pedestrians in a crowded scene is constructed, with the edge weights indicating how much overlap and interaction between these bounding boxes exists.

Let $A_i \subset \mathbb{N}^2$ with $i \in \mathcal{I}$ be a set of pixel coordinates comprised by the persons corresponding bounding box. The generalized Crowd Ratio $\xi^{(i,k)} : \mathbb{N}^2 \times \mathbb{N}^2 \to [0,1]$ is then defined as the harmonic mean

$$\xi^{(i,k)} := \frac{2 \cdot \psi_{ik} \cdot J_{ik}}{\psi_{ik} + J_{ik}} \tag{4.1}$$

between two functions $\psi : \mathbb{N}^2 \times \mathbb{N}^2 \to [0,1]$ with

$$\psi(A_i, A_k) := \frac{\min\{|A_i|, |A_k|\}}{\max\{|A_i|, |A_k|\}} =: \psi_{ik} \tag{4.2}$$

and the Jaccard[1] index $J : \mathbb{N}^2 \times \mathbb{N}^2 \to [0,1]$ with

$$J(A_i, A_k) := \frac{|A_i \cap A_k|}{|A_i \cup A_k|} =: J_{ik} \tag{4.3}$$

respectively.

Now the connection graph is an undirected, weighted graph $G := (V, E, \Xi)$ with the pedestrians as nodes ($V := \mathcal{I}$) and edge weights $\xi^{(i,j)} = \xi^{(j,i)} \in \Xi = (\xi^{(i,j)})_{0 \le i,j \le n-1}$, and $E := \{(i,j) \mid \xi^{(i,j)} > 0\}$.

In order to capture the overall crowdedness of a given scene, ψ and J aim to rate visually interacting people. Thereby, J measures the actual closeness between two arbitrary people, whereas ψ rates this closeness by taking the actual size of the respected people into account. In other words, it aims to generate higher scores for similar sized people, i.e., those standing side-by-side, and lower scores for those strongly differing with respect to size. The latter is most often the case when one person is much closer to the camera than the other. Figure 4.6 visualizes two exemplary bounding boxes A_i and A_k, with the areas $|A_i| = a_0 \cdot a_1$, $|A_k| = b_0 \cdot b_1$, and the resulting intersection $X = A_i \cap A_k$ with $|X| = x_0 \cdot x_1$.

[1] The Jaccard index is also known as the Intersection over Union (IoU) in computer vision.

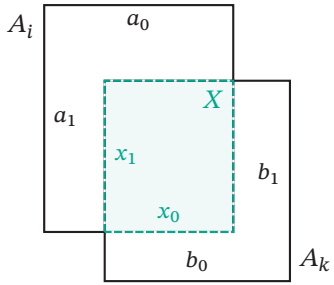


Figure 4.6: Geometrical visualization of two intersecting bounding boxes A_i and A_k. The intersection area X is indicated by the colored area.

The GCI as presented above is already suited for characterizing crowded scenarios, however since the measure is based on areas rather then on single points, the values for the GCI are in general far smaller compared to the CI. This is mainly due to the fact that the areas $|A_i|$ are larger then the number of keypoints, especially for pedestrians closer to the camera, which is emphasized by the choice of the harmonic mean. The more J_{ik} and ψ_{ik} diverge from each other, the lower the score is, meaning that people that move away from each other will rapidly generate a lower score. As mentioned earlier both, the original and the presented metric, are computed the same way. Furthermore, the CR as introduced for the CI can be rewritten in such way that the resulting score can be determined given the same formula as shown in Equation (4.4).

$$\xi_i = \sum_{k=0}^{n-1} \xi^{(i,j)} \tag{4.4}$$

In order to achieve this, each term of the sum of the CR (see Equation (3.4)) has to be separated into different terms $\xi^{(i,j)}$ so that pairwise scores are obtained. This results in $\xi^{(i,k)} := \frac{n_i^k}{n_i^i}$ for every element of the sum. Note that for both, the CI and GCI $\xi^{(i,k)}$ is omitted if $i = k$ since the comparison of a person with itself is unnecessary. The resulting formula for the CI and GCI looks

than as follows:

$$f = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} \xi^{(i,k)} \tag{4.5}$$

Rearranging the single terms of the double sum results then in the following formula:

$$\begin{aligned} f &= \frac{1}{n} \sum_{i=0}^{n-1} \left(\xi^{(i,i)} + \sum_{k=0}^{n-1} \xi^{(i,k)} + \sum_{k=0}^{n-1} \xi^{(k,i)} \right) \\ &= \frac{1}{n} \sum_{i=0}^{n-1} \left(\xi^{(i,i)} + \sum_{k=i+1}^{n-1} \underbrace{\left(\xi^{(i,k)} + \xi^{(k,i)} \right)}_{=:r} \right) \end{aligned} \tag{4.6}$$

As mentioned above, $\xi^{(i,i)}$ are excluded, which can be mathematically expressed by $\xi^{(i,i)} := 0 \; \forall i \in \{0, \dots, n-1\}$. After bringing the formula to the form in Equation (4.6), the comparison between both metrics comes down to the non-diagonal elements of Ξ that are part of the reminding term. To be more concise, the non-diagonal elements, i.e., $\xi_{\mathrm{CI}}^{(i,k)}$ and $\xi_{\mathrm{GCI}}^{(i,k)}$ with $i,k \in \{0, \dots, n-1\}$ and $i \neq k$, can fall into different cases such as:

Case 1 The bounding boxes have no intersection. From the definition of the CR for CI and GCI follows directly $\xi_{\mathrm{CI}}^{(i,k)} = \xi_{\mathrm{GCI}}^{(i,k)} = 0$.

Case 2 The bounding boxes intersect, but the intersection does not contain any keypoints. Hence, since there is an intersection and $J_{ik} > 0$ and $\psi_{ik} > 0$. Consequently, $\xi_{\mathrm{CI}}^{(i,k)} = 0$ and $\xi_{\mathrm{GCI}}^{(i,k)} > 0$.

Case 3 Intersection containing at least one keypoint. Assumption $\min\{ \xi_{\mathrm{CI}}^{(i,k)}, \xi_{\mathrm{CI}}^{(k,i)} \} - \epsilon \leq \xi_{\mathrm{GCI}}^{(i,k)} \leq \max\{ \xi_{\mathrm{CI}}^{(i,k)}, \xi_{\mathrm{CI}}^{(k,i)} + \epsilon \}$; One trivial solution would be $\epsilon = 1$, since it defines the smallest threshold that covers the whole codomain of ξ.

These cases indicate that there is no obvious relation between the two metrics. Furthermore, due to the fact that $\xi_{\mathrm{CI}}^{(i,k)}$ is a discrete and $\xi_{\mathrm{GCI}}^{(i,k)}$ is a continuous

Table 4.3: Statistics computed on the SyMPose dataset. The population size is $N = 20{,}808$ based on 100 randomly drawn frames from the dataset. Since $\xi_{\text{CI}}^{(i,k)}$ takes only certain values, for each value of $\xi_{\text{CI}}^{(i,k)}$ the mean, and the 95% and 99% prediction intervals $h_{0.95}$ and $h_{0.99}$ of $\xi_{\text{GCI}}^{(i,k)}$ are shown.

	$\xi_{\text{GCI}}^{(i,k)}$		
$\xi_{\text{CI}}^{(i,k)}$	μ	$h_{0.95}$	$h_{0.99}$
0.000000	0.0515	0.0015	0.0019
0.071429	0.1284	0.0027	0.0036
0.142857	0.1755	0.0033	0.0043
0.214286	0.2350	0.0036	0.0047
0.285714	0.2722	0.0042	0.0056
0.357143	0.3290	0.0051	0.0068
0.428571	0.3715	0.0050	0.0065
0.500000	0.4072	0.0058	0.0077
0.571429	0.4479	0.0059	0.0077
0.642857	0.4982	0.0075	0.0099
0.714286	0.5235	0.0070	0.0092
0.785714	0.5843	0.0084	0.0111
0.857143	0.6264	0.0089	0.0117
0.928571	0.6235	0.0111	0.0146
1.000000	0.6714	0.0165	0.0217

function, a complete analytical comparison is even trickier. However, from an empirical perspective both metrics can be easily collated as it is done in the following. Observing the difference between the original CI and the generalized GCI, shows that in general the latter produces lower values than the CI. This is underlined by the data shown in Table 4.3, which outlines that for very small values of $\xi_{\text{CI}}^{(i,k)}$, i.e., only few foreign keypoints of a person k interfering with the current person i, $\xi_{\text{GCI}}^{(i,k)}$ is generally larger, but with increasing overlap $\xi_{\text{CI}}^{(i,k)}$ tends to yield larger values than $\xi_{\text{GCI}}^{(i,k)}$. Figure 4.7 illustrates various cases with three examples taken from the SyMPose dataset.

(a)

$\xi_{\text{CI}}^{(i,k)} = 0.0714$
$\xi_{\text{CI}}^{(k,i)} = 0.4286$
$\xi_{\text{GCI}}^{(i,k)} = 0.3330$

(b)

$\xi_{\text{CI}}^{(i,k)} = 0.7143$
$\xi_{\text{CI}}^{(k,i)} = 0.1429$
$\xi_{\text{GCI}}^{(i,k)} = 0.3381$

(c)

$\xi_{\text{CI}}^{(i,k)} = 0.7143$
$\xi_{\text{CI}}^{(k,i)} = 0.9286$
$\xi_{\text{GCI}}^{(i,k)} = 0.7158$

Figure 4.7: Exemplary cases for $\xi_{\text{CI}}^{(i,k)}$ and $\xi_{\text{GCI}}^{(i,k)}$. Here, person i in yellow is adduced for comparison. Therefore, (a) gives an example where the GCI has the biggest negative difference, i.e., is larger than CI, whereas in (b) the GCI shows the biggest difference in the opposite direction, i.e., GCI is smaller than CI, and finally (c) visualizes an example where the GCI has closest absolute distance from the CI. For better understanding, person i is shown in yellow and person k in red.

(a)

$\bar{\xi}_{\text{CI}}^{(i,k)} = 0.1429$
$\xi_{\text{GCI}}^{(i,k)} = 0.1412$

(b)

$\bar{\xi}_{\text{CI}}^{(i,k)} = 0.7857$
$\xi_{\text{GCI}}^{(i,k)} = 0.5508$

Figure 4.8: Exemplary cases for a symmetrical interpretation of the CI. The figure follows the visualization given in Figure 4.7. Here, $\bar{\xi}_{\text{CI}}^{(i,k)} = \bar{\xi}_{\text{CI}}^{(k,i)} = \frac{1}{2} \cdot (\xi_{\text{CI}}^{(i,k)} + \xi_{\text{CI}}^{(k,i)})$. On the left, (a) shows the lowest score, with both boxes showing similar posture resulting in similar boxes and small overlap, and on the right in (b) the largest difference between the two scores, resulting by a high overlap, with a strong deviance between the two postures.

All of these examples are typical situations that fall into the third case, i.e., where the intersection between two pedestrians contains at least one keypoint. Figure 4.8 extends these with a minor change in the way $\xi_{\text{CI}}^{(i,k)}$ is computed, namely by making Ξ symmetrical. This is achieved by averaging

its non-diagonal corresponding elements, i.e., $\xi_{\mathrm{CI}}^{(i,k)}$ and $\xi_{\mathrm{CI}}^{(k,i)}$, which yields $\bar{\xi}_{\mathrm{CI}}^{(i,k)} = \bar{\xi}_{\mathrm{CI}}^{(i,k)} = \frac{1}{2} \cdot (\xi_{\mathrm{CI}}^{(i,k)} + \xi_{\mathrm{CI}}^{(k,i)})$. Following this approach does not change the overall score but reduces the variance for the $\xi_{\mathrm{CI}}^{(i,k)}$ since the values for interfering pedestrians are often far apart from each other.

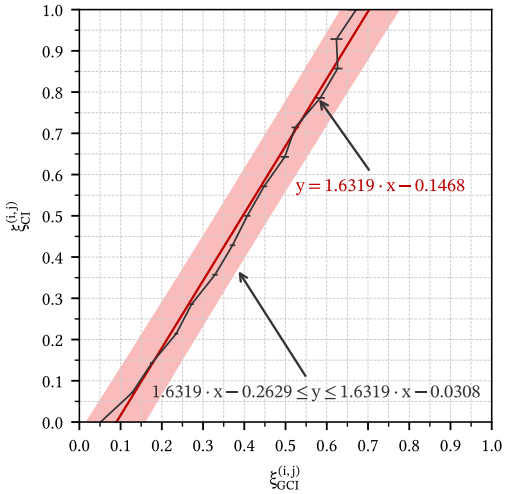


Figure 4.9: Results of a linear regression based on 100 random frames ($N = 20{,}808$ pedestrian samples) from the SyMPose dataset. The black curve shows the mean and deviation of $\xi_{\mathrm{GCI}}^{(i,k)}$ resulting for different values of $\xi_{\mathrm{CI}}^{(i,k)}$ signifying an approximately linear relation between both functions. The red line therefore is the result of the aforementioned linear regression. Furthermore, the gray shaded area indicates the prediction band that contains 99% of all samples. Note that $x \coloneqq \xi_{\mathrm{GCI}}^{(i,k)}$ and $y \coloneqq \xi_{\mathrm{CI}}^{(i,k)}$ for a clearer visualization.

As Table 4.3 already indicated, with increasing overlap between two arbitrary pedestrians i and k, i.e., CI and GCI increasing, the ratio $\xi_{\mathrm{GCI}}^{(i,k)} : \xi_{\mathrm{CI}}^{(i,k)}$ drops below 1. Figure 4.9 visualizes the data from Table 4.3 and illustrates the observed approximate linear relation between both scores. Based on a linear regression this link can be described and a prediction band derived that allows for an estimated upper and lower boundary for $\xi_{\mathrm{GCI}}^{(i,k)}$ in which 99% of

the samples fall into. In this particular sample population, $\epsilon = 0.1160$ which bounds $\xi_{\text{GCI}}^{(i,k)}$ approximately by

$$f(\xi_{\text{GCI}}^{(i,k)}) - 0.1160 \leq \xi_{\text{CI}}^{(i,k)} \leq f(\xi_{\text{GCI}}^{(i,k)}) + 0.1160 \tag{4.7}$$

with $f(x) \coloneqq y$. So far, the GCI successfully generalizes the CI and extends it for surveillance scenarios. However, despite the changes already made, the GCI in its pure form fails to meet the expectation for certain constellations in such crowded situations.

In the following, two members of the GCI family will be introduced that aim to improve the basic idea presented above. Their main motivation is an appropriate reweighing of single CRs, which covers in particular people that take up larger spaces due to their posture.

4.2.3.2 Normalized Graph Crowd Index

The first of two proposed extensions of the general Graph Crowd Index idea is the so-called Normalized Graph Crowd Index (nGCI). Therefore, in a first step, a centrality score based on the connection graph is defined for each pedestrian i as

$$c_i \coloneqq \tilde{\sigma}\Big(\sum_{j=0}^{n-1} \xi^{(i,j)}\Big) \tag{4.8}$$

where $\tilde{\sigma} : \mathbb{R} \to (-1,1) \subset \mathbb{R}$ is a modified version of the logistic function which controls how large values are handled. In this case it is chosen as

$$\tilde{\sigma}(x) \coloneqq \frac{2}{1 + e^{-x}} - 1 \tag{4.9}$$

Secondly, a scale weight for each pedestrian i is defined as

$$s_i \coloneqq \sqrt{|A_i|} \tag{4.10}$$

Now, the final result that is returned is the weighted average

$$\text{nGCI} := L \cdot \frac{\sum_{i=0}^{n-1} c_i \cdot s_i}{\sum_{i=0}^{n-1} s_i} \tag{4.11}$$

where L is a scale constant which is empirically set to 5 to make the nGCI comparable to the CI.

4.2.3.3 Standard Graph Crowd Index

The sGCI differs from CI and nGCI to the extent that the obtained score is not normalized by the number of pedestrians in any way. Hence, with more pedestrians in the scene, the score will always get higher. The centrality scores c_i for the pedestrians are calculated in the same way as for the nGCI. The scale weights are

$$s_i := \frac{\sqrt{|A_i|}}{\frac{1}{n}\sum_{j=0}^{n-1}\sqrt{|A_j|}} \tag{4.12}$$

and now the final result is

$$\text{sGCI} := L \cdot \sum_{i=0}^{n-1} c_i \cdot s_i \tag{4.13}$$

where L is a scale constant which is set to 0.05 to make the sGCI comparable to CI.

Both, the sGCI and nGCI are plotted with respect to the original CI in Figure 4.10. The figure shows, that with respect to the categorization both measures show a similar outcome as the original CI. The sGCI in particular brings the scores closer to the lower end for the majority of samples.

For more information on the CI as well as the GCI, see Appendix A.3.

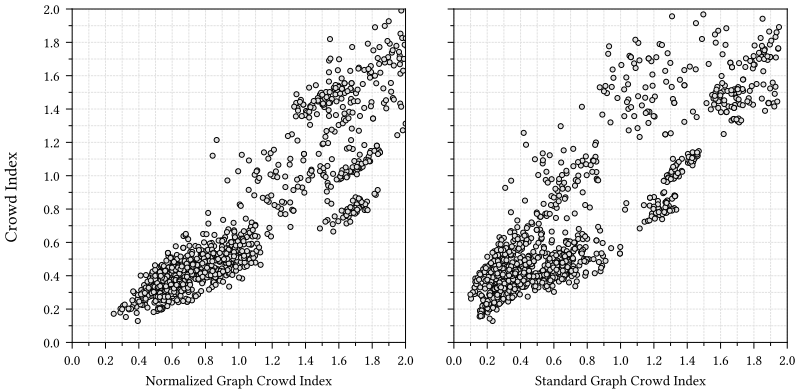


Figure 4.10: Comparison of the two introduced metrics nGCI and sGCI with the CI on sampled frames from the synthetic SyMPose dataset. The chosen configurations shows that both metrics are appropriate to replace the CI, especially since there is no need for keypoint annotations.

4.2.3.4 Summary

This section proposed a new family of metrics for quantifying the crowdedness of a scene. Inspired by the CI, a generalized approach was presented that attempts to implement the properties of the CI for scenarios and datasets that do not provide keypoint information. Furthermore, two members of the GCI family were introduced bringing the general GCI closer to the value range of the CI and emphasizing properties of typical video surveillance views such as the lower variance in sizes and larger number of pedestrians.

For the remaining part of this thesis, the metric of choice is the sGCI as it is more general than the CI. From empirical observations the sGCI has shown its suitability for the kind of data this work focuses on. Both aspects underline this decision.

4.3 Style Transfer for Domain Adaptation

The preceding section gave an introduction to one way that allows for privacy-friendliness for the task of human behavior analysis by taking a look at the topic of how to acquire suitable training data. Its main idea as presented in Section 4.2 is tailoring custom training data to the corresponding task, in particular the development of skeleton-related methods in a broader sense. Generally speaking, relying on synthetic data brings up various challenges, like a limited set of behaviors and movements, as well as a certain visual bias. Therefore, the following pages address the issue of differing appearance between synthetic and real-world data by applying methods for DA, which is a particular form of transfer learning (TL). In particular the section starts by introducing the task of DA, followed by some words on challenges that come up when trying to adapt data for surveillance scenarios, motivating the necessity of expansions. This is followed by the presentation of a DA approach for video surveillance scenarios showing crowded scenes, which expands the methods introduced in Section 3.3. Finally, the real-world dataset is presented, which will take on the role of the target domain in the experiments, completing the pair of source and target domain.

4.3.1 Domain Adaptation

The aim of DA is to leverage existing labeled data from one or more related source domains as training data for a supervised machine learning algorithm that is applied to unseen or unlabeled data in a target domain. Both domains are assumed to be somehow related but not similar [Csu17b]. The prospect of providing such a machine learning model has made DA a heavily investigated field, especially in applications where the cost of gathering labeled data is high. However, training a model in the source domain without taking into account the dissimilarity between the two domains nearly always results in performance reduction when testing or applying the model in the target domain [Csu17b]. The amount of degradation in performance depends on how close the domains are related.

More formally, a domain $\mathcal{D}$ is composed of a d-dimensional feature space $\mathcal{X} \in \mathbb{R}^d$ with a marginal probability distribution $p(\boldsymbol{X})$, and a task $\mathcal{T}$, which is defined by a label space $\mathcal{Y}$ and a conditional probability distribution $p(\boldsymbol{Y}|\boldsymbol{X})$, with random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ [Pan10, Wei16b, Csu17b]. Typically, supervised machine learning models are trained to infer $p(\boldsymbol{Y}|\boldsymbol{X})$ from a dataset consisting of samples $\{\mathbf{x}_0, ..., \mathbf{x}_{N-1}\}$ drawn from $p(\boldsymbol{X})$ and corresponding labels $\{y_0, ..., y_{N-1}\}$ drawn from $p(\boldsymbol{Y})$. However, in the case of a TL scenario, there are two domains as stated above:

- the source domain $\mathcal{D}^s = \{\mathcal{X}^s, p(\boldsymbol{X}_s)\}$ with a task $\mathcal{T}^s = \{\mathcal{Y}^s, p(\boldsymbol{Y}_s|\boldsymbol{X}_s)\}$, and
- the target domain $\mathcal{D}^t = \{\mathcal{X}^t, p(\boldsymbol{X}_t)\}$ with a task $\mathcal{T}^t = \{\mathcal{Y}^t, p(\boldsymbol{Y}_t|\boldsymbol{X}_t)\}$ [Csu17b]

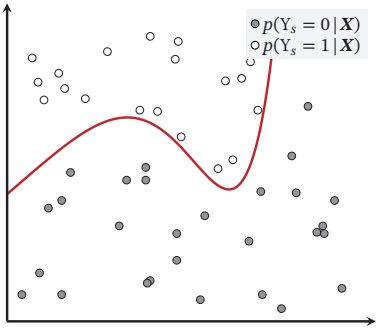
For the scope of this thesis, the scenario of interest is a particular case where the feature and label spaces of the two domains equal

$$\mathcal{X}^s = \mathcal{X}^t \,, \tag{4.14}$$

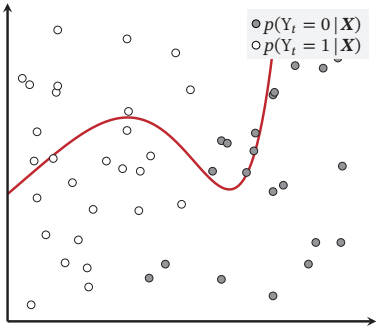
$$\mathcal{Y}^s = \mathcal{Y}^t \,, \tag{4.15}$$

but the marginal probabilities are assumed to be different, $p(\boldsymbol{X}_s) \neq p(\boldsymbol{X}_t)$. This inequality is commonly also referred to as *domain shift* and causes the conditional probabilities $p(\boldsymbol{Y}_s|\boldsymbol{X}_s)$ and $p(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ to differ from each other and is the main reason for the performance decrease depicted above.

Figure 4.11 illustrates a common issue that arises from domain shift for a binary classification task: The classifier's decision boundary, which is marked in red, is learned in the source domain on instances of which are drawn from the marginal distribution $p(\boldsymbol{X}_s)$. As the marginal distributions of the source and target domains differ from each other, i.e., $p(\boldsymbol{X}_t) \neq p(\boldsymbol{X}_s)$, the decision boundary fails to separate the two classes in the target domain appropriately. Here come DA methods into the picture, as they aim at solving this issue. According to the definitions of Pan et al. [Pan10], DA can be categorized as *homogeneous, transductive TL*, where *homogeneous* refers to Equation (4.14) and *transductive* refers to Equation (4.15). Pan et al. [Pan10] provide a broader



(a) Illustration of a binary classifier's decision boundary, that is learned from samples drawn from $p(\boldsymbol{X}_S)$.



(b) Failure of the classifier when applied samples drawn from $p(\boldsymbol{X}_t)$.

Figure 4.11: Failure of binary classifier caused by domain shift between the source and target domain: its decision boundary is learned in the source domain $\mathcal{D}^s$, as shown left. When being applied in the target domain $\mathcal{D}^t$, the classifier fails in separating the two classes as the examples are drawn from the marginal distribution $p(\boldsymbol{X}_t)$ that differs from the distribution $p(\boldsymbol{X}_S)$, which the training examples were drawn from. This is depicted on the right side [Bla19].

overview on this topic and gives a complete categorization of various subdomains of TL. Moreover, it can be distinguished between the *unsupervised* case, where the labels are only available in the source domain, and the *supervised* case, where at least a small number of labeled instances is also provided in the target domain.

As the goal described in the beginning of this section is to obtain suited training data for the target domain due to the lack of annotations for such, in this work, labels are only assumed to be available for the synthetic data, which serves as source domain and will subsequently be denoted as $\mathcal{D}^s$. The real-world data which depicts the target domain $\mathcal{D}^t$ is assumed to be entirely unlabeled with respect to the task of HPE. Therefore, only the case of *unsupervised DA* is regarded within this thesis.

(a) Original **(b)** Adapted

Figure 4.12: A single frame from SyMPose and its adapted counterpart. The DA fails to preserve small pedestrians, which consequently have to be filtered out [Bla19].

4.3.2 Challenges for Video Surveillance Applications

Certain properties of video sequences from surveillance cameras cause an increase in difficulty to a GAN-based DA framework. The preservation of content between the original and the adapted image is crucial when intending to use the adapted images as training samples. In the case of HPE, this is especially important for the persons that an image contains: all people that are shown within the image have to be preserved between the domains. Otherwise, there would be keypoints included in the annotations, that do not correspond to a person in the adapted image. The result would be low-quality data so that a model trained on that data tends to perform poorly on real-world images. In the special case of frames that are obtained from video surveillance cameras in urban settings, there exist many persons at small scales. This comes from the cameras being mounted in elevated positions with a potentially wide view. Therefore, current state-of-the-art DA methods, such as the ones introduced in Section 3.3, are likely to fail to preserve all people as they solely conserve the dominant parts of a scene. Considering the situation shown in Figure 4.12, which shows a single frame before and after adaptation by the Cycle-GAN, it becomes obvious that only the persons at larger scale are still well recognizable after the domain translation. In this case the model was trained to adapt synthetic images from the SyMPose dataset to the Cityscapes [Cor16] dataset [Bla19]. Using such images for training would require a manual filtering of the existing annotations, which is a step that is

intended to be avoided. This emphasizes the need of expansions to ensure the preservation of all persons.

Furthermore, depending on the scenario of a surveillance setup, various visual influences increase the task's difficulty, especially in outdoor settings that are strongly affected by the current weather. This for instance can lead to strongly differing visual appearances. Interfering rain drops on the weather shield of a camera can have various effects on the recorded image, but also the white balance of the camera that has to compensate for the changes in illumination has direct influence on the overall appearance of the scene. Finding suited datasets that can serve as target domain is difficult as such datasets should also exhibit at least a small portion of shared properties between its images.

Thus, the required stabilization of the GAN optimization is challenging for such a situation. In addition to the experiments on public datasets [Bla19, Gol20] like the WorldExpo '10 dataset [Zha15] this thesis provides experiments on a custom dataset. It fits the domain of the HPE task, satisfies the properties that were mentioned earlier and comes with a smaller domain shift to the source domain, which is defined by the SyMPose dataset. The created target data is presented in Section 4.3.5.

4.3.3 Cycle-GAN Approach

As indicated earlier, the goal is to alter the appearance of the existing synthetic data in such way, that when used for training of an HPE method, the overall performance increases. In other words, the proposed approach aims on closing the existing domain gap. This is referred to as *style transfer* [Bla19, Gol20] and is mostly known for those approaches that alter the style of pictures to look like they have been drawn by certain artists like van Gogh, Monet or Picasso [Gat15]. Different to the art setup, detailed structures are important to be maintained as they carry information in particular for visible pedestrians. The studies in [Bla19, Gol20] investigated different ways to achieve such style transfer using different kinds of GAN-based models and their extensions.

The core architecture chosen for the style transfer part of this thesis is the so-called Cycle-GAN introduced by Zhu et al. [Zhu17], which has been presented briefly in Section 3.3.1. It has been applied to many tasks, such as unpaired image-to-image translation [Lon19, Zha20] and neural style transfer [Cur20, Gao20, She21]. The DA or style transfer task requires the generated sample $\tilde{\mathbf{x}}$ to preserve the content of the input image $\mathbf{x}_s$. Therefore, the general GAN objective $\mathcal{L}_G$ as presented in Section 2.1.5, which only ensures G solely to meet the target distribution $p(\boldsymbol{X}_t)$, is not sufficient and has to be exchanged by a more restrictive formulation. The so-called *cycle-consistency loss* is given in Equation (4.16) and attempts to reduce the degrees of freedom in order to achieve the mentioned goal.

$$\begin{aligned}\min_{G_s,G_t} \mathcal{L}_{\text{cyc}}(G_s, G_t, \boldsymbol{X}_s, \boldsymbol{X}_t) = \mathbb{E}_{\mathbf{x}_s \sim p(\boldsymbol{X}_s)}\Big[\|G_t(G_s(\mathbf{x}_s)) - \mathbf{x}_s\|_1\Big] \\ + \mathbb{E}_{\mathbf{x}_t \sim p(\boldsymbol{X}_t)}\Big[\|G_s(G_t(\mathbf{x}_t)) - \mathbf{x}_t\|_1\Big]\end{aligned} \tag{4.16}$$

The Cycle-GAN is realized using a U-Net architecture for the generator part as well as a patch discriminator. U-Net is a widely known encoder-decoder-network that was firstly applied for the segmentation of neuronal structures [Ron15]. As it captures and preserves context information, it has also been utilized in various GAN-based image-to-image-translation frameworks such as [Iso17]. The central idea is to concatenate every layer's output during the down-sampling in the encoder network with the corresponding up-sampling layer output, which has the same spatial dimensions in the decoder network and thus ensuring a transport of extracted information from encoder to decoder part. Table 4.4 shows the resulting network architecture of the U-Net-generator as used in this work. Given an input image $\mathbf{x}_s$, multi-level feature information is extracted by the encoder network and directly transmitted to the corresponding decoder layer via skip connections. Thus, content information have a stronger impact on the adapted image $\tilde{\mathbf{x}}$ than for the structure without the skip connections.

Table 4.4: Architecture of the U-Net generator as presented in [Bla19, Gol20] consisting out of various layer types. *Conv* layers are convolutional layers followed by instance normalization and ReLU activation. *Res* layers are residual blocks [San18] consisting of two convolutional layers, each followed by instance normalization and a final ReLU activation. Finally, *deconv*-layers are convolutional layers followed by instance normalization and up-sampling layer.

	type	kernel size	stride	filters
encoder	conv	7	1	32
	conv	3	2	64
	conv	3	2	128
	conv	3	2	256
residual	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
	res	-	-	256
decoder	deconv	3	2	256
	deconv	3	2	128
	deconv	3	2	64
	conv	3	1	3

4.3.4 Training Aspects

The U-Net-based Cycle-GAN introduced so far addresses the first of the challenges presented earlier. In order to decrease the influence of disruptive dataset properties, the following data augmentation techniques are applied to image samples from the source as well as from the target domain:

- Reflection padding at the image borders with a quarter of the image size; this is done to avoid zeros at the borders of the augmented images
- Random cropping to a fixed training image size, without scaling
- Random image rotations within a range of [-20°, 20°] with a probability of 0.5

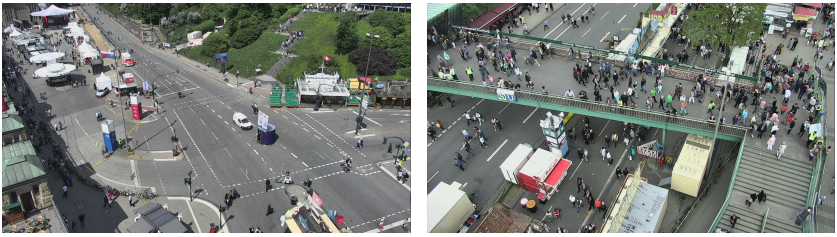
- Random horizontal and vertical flips with probabilities of 0.8 and 0.9
- Random color changes in the HSV-color-space for the *hue* and *saturation* channels: The *hue* angle changes within the range [-9°,9°]; the saturation value is multiplied with a factor inside the range [0.95,1.05] and clipped within [0,1]; both modifications are carried out with a probability of 0.8

After these operations, the images are normalized in between [-1,1], which is a common form of normalization for GANs, since the generator's final activation is the hyperbolic tangent function. These data augmentations, which are rather unusual for typical GAN models, help limiting the discriminator's power by ensuring that well recognizable image components do not appear in every sample. As the content shall be conserved anyhow, it is not an issue to the generator, if images are flipped or rotated. Blattmann [Bla19] showed with respect to the Cityscapes dataset, that the manufacturer emblem of the car is an illustrative example for such an image component. It is contained in nearly every image at the approximately same region and could thus be easily recognized by the discriminator, if it was present in every training example. As a consequence, the generator would learn to put this emblem into the adapted images, too, what should be avoided for obvious reasons. A similar situation comes up when regarding static cameras in a surveillance setup with certain architectural structures that exist in every frame. As a result of the cropping and rotation operations, the discriminator cannot remember image artifacts like this that well and thus, there is no need for the generator to produce images showing that specific sign. Figure 4.13 shows augmented images for all the target domain dataset.

4.3.5 Target Domain Dataset

Selecting suited data as target domain is crucial as shown in [Bla19, Gol20]. As Blattmann [Bla19] stated an eligible dataset should exhibit at least a small portion of shared properties between its images. For instance, the illumination conditions should be approximately equal for all images in a valid target domain dataset. If such common properties are missing, the DA is likely to fail

Figure 4.13: Exemplary results of the data augmentation process. The atypical strong spatial variations are mainly performed to limit the power of the discriminator.



(a) Camera 1 **(b)** Camera 3

Figure 4.14: Two examples from DA target dataset recorded during Hafengeburtstag Hamburg in 2018. Most of the dataset was recorded during daytime showing a wide range of pedestrian counts from few single ones in the early hours up to full streets with hundreds of pedestrians.

because of lacking characteristics of the target domain, that the generator can identify. A possible solution to this is to use data from a single video sequence or a single statically mounted surveillance camera. This, however, causes the discriminator network to become too powerful as the categorization of real and adapted images does not need too much effort. The reason for this is a great amount of shared image content between the images in such a dataset, for instance static objects, like a parked car, that are contained in every image. Given the knowledge about the application domain and the finding by Blattmann [Bla19], a dataset was created based on data collected during Hafengeburtstag Hamburg in 2018 which shall be used as target dataset for DA in video surveillance. Figure 4.14 shows two examples from the dataset The dataset is founded on recordings from May 11, 2018. On this day, 245 videos were recorded throughout the day around 8 a.m. to 11 p.m. The videos

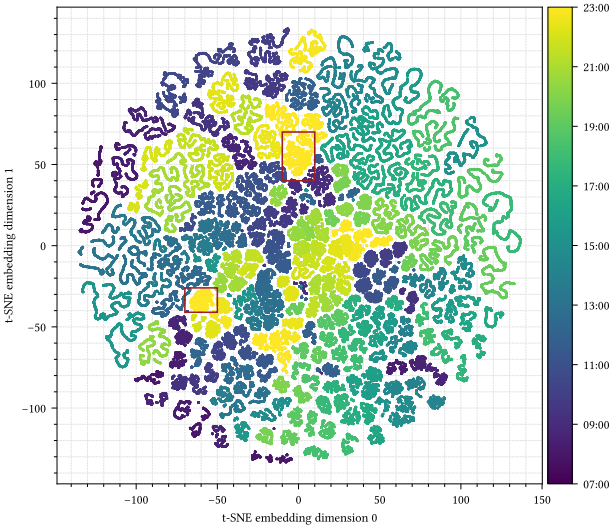


Figure 4.15: Resulting t-SNE dimensionality reduction based on image features extracted using VGG16. As indicated by colorbar on the right, the colors correspond to the time the videos were recorded. Note: This figure illustrates a subset of all feature points due to computational cost in visualizing.

were captured by four identical Sony SNC-WR632C[1] cameras positioned at the same height on a tower, each pointing in a different direction. In order to establish the data basis, 500 key frames were randomly chosen from each of the 245 videos, resulting in 122,500 single images in the dataset. The t-Distributed Stochastic Neighbor Embedding (t-SNE) [Maa08] dimensionality reduction algorithm was used to illustrate the characteristics of the dataset based on features retrieved by the VGG16 [Sim14] feature extractor. The visualization shows, that all points could be matched to certain neighborhoods, however several data points show some randomness, lacking typical structures and patterns as the remaining data. A closer look on these samples, highlighted with an red rectangle shows that all of the data points are rather noisy due to the time they were recorded and the resulting noisy image. This

[1] For more information see https://pro.sony/en_GB/products/ptz-security-cameras/snc-wr632c

observation was made throughout various t-SNE runs with different levels of perplexity ranging from 0 to 30 in steps of 5. For a better understanding of the visual appearance of the target domain dataset see Figure A.9 in Appendix A.5.1.

4.4 Motion Analysis of Pedestrians

The third part of this thesis copes with algorithms and methods to evaluate human behavior based on motion analysis given the urban surveillance setting. Again, the approaches presented in the course of this section aim to be privacy-friendly, which is mainly achieved by the choice of input features.

Therefore, this section starts by giving a definition of the actual problem space. Based on this definition two different approaches are introduced: a holistic pedestrian-agnostic method using general motion information extracted from videos, and a human-centric approach that explicitly focuses on a temporal view on structural information of humans.

4.4.1 Definition of the Problem Space

In general, human behavior is very complex and condensing the set of every possible behavior to few classes can be very demanding. Furthermore, the behavior of individuals strongly depends on the context and the given situation, i.e., certain behavior that might be unexpected in one type of situation, can be completely normal in other circumstances. From the authorities' perspective, most of the conceivable behaviors, denoted as X, are not of interest since they pose no threat to other pedestrians. But it is not just the behavior that directly exposes threat to others that draws the attention of security staff and police. Lurching pedestrians and mentally disturbed beings are just two examples of situations where people are a danger to themselves. Yet there exist certain behaviors that are typically characterized as aggressive or harmful, and

draw the interest of authorities responsible for public security and safety. Figure 4.16 illustrates in an abstract way that the relevant behaviors that call for intervention constitute only a small subset of the set of possible behaviors X.

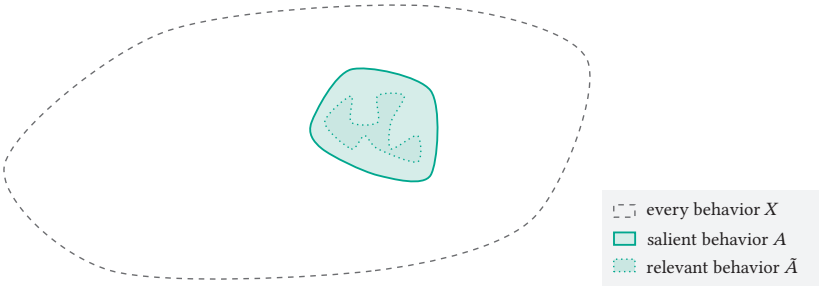


Figure 4.16: This schematic illustration depicts various categories of behavior. Salient and relevant behaviors are naturally encompassed within the broader set of behaviors, denoted as X. The potentially complex subset representing relevant behaviors, $\tilde{A}$, is contained within its relaxation set A, also referred to as *salient behavior* throughout this thesis.

Thinking about this subset in a mathematical way, denoted as $\tilde{A}$, might even exhibit a complex structure, which possibly can consist of non-connected subsets due to the heterogeneity of the included behaviors. Such complex structure makes it difficult to understand its elements and their relation, so that it might be not enough to have one function to implement $\mathbf{1} : X \to \{0, 1\}$ as the indicator function, with

$$\mathbf{1}(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{x} \in \tilde{A} \\ 0 & \text{else} \end{cases} \tag{4.17}$$

To simplify the target space, a more general set A with $\tilde{A} \subset A$ is defined. This set can be regarded as a generalization of $\tilde{A}$ and serves as an approximate model of the set of relevant behaviors. Consequently, the hope is to find or learn a function $f(\mathbf{x})$ with $f(\mathbf{x}) \approx \mathbf{1}_R(\mathbf{x})$, which can be seen as the relaxation

of $\mathbf{1}(\mathbf{x})$ and

$$\mathbf{1}_R(\mathbf{x}) \coloneqq \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{else} \end{cases} \tag{4.18}$$

Furthermore, if $|X| \gg |\tilde{A}|$, then $|X| \gg |A|$ should hold as well, which implies that A ideally holds as many elements as necessary and still be easy to learn. One potential advantage of choosing A is that it encompasses even those behaviors that authorities may not have considered previously. In other words, A contains possibly aggressive behaviors, sudden behaviors that seem harmful or dangerous in the first place, and further behaviors differing from common or expected behaviors (i.e., anomalies). For a better understanding, Figure 4.17 illustrates a simplified example situation of two similar dynamics with different levels of threat. The two protagonists, a female and a male, are part of two situations that show a very similar dynamic throughout the video, especially by having the same beginning where the men is running towards the woman who is leaving her car. As the figure shows, one video is actually the male attacking the female, whereas the other is exaggerated greeting. In both cases an alarm would have been appropriate, especially since both show an increasing threat level from start of the video to the encounter. Note that apart from the overall dynamic, the outfit of the male protagonist has obviously an additional influence on the way the viewer might classify the situation. This is an issue not regarded within this thesis.

About Anomalies and Abnormalities

In the field of computer vision and other technical fields, salient instances are typically referred to as anomalies. However, from a linguistic perspective, there exists another term, which is sometimes used interchangeably, namely *abnormality* or *abnormalities*. The Oxford dictionary refers to an *anomaly* as "a thing, situation, etc. that is different from what is normal or expected", whereas an *abnormality* is "a feature or characteristic in a person's body or behaviour that is not usual and may be harmful or cause illness or worry; the fact of having such a feature or characteristic". In other words, an anomaly

Figure 4.17: This schematic illustration shows two cases of behavior. With respect to the male person in the shown frames, the initial situation is identical as the man is running towards the woman. Although, there is nothing wrong with the man running, it results in two different outcomes, where one is relevant (top) and the other not (bottom). Both situations are difficult to distinguish in an automatic manner and even for a human operator an early information on both situations would be beneficial.

refers to any instance or pattern within a dataset that deviates significantly from the norm or expected behavior. Anomalies are typically identified as data points that are rare, unexpected, or inconsistent with the majority of the data. They can be indicative of errors, outliers, or interesting insights depending on the context. An abnormality, on the other hand, is a broader term that encompasses deviations from what is considered normal or typical within a dataset. While anomalies are a subset of abnormal data, abnormality can include a wider range of deviations, such as outliers, errors, inconsistencies, or simply data points that fall outside the norm. Abnormality can be subjective and context-dependent, as what is considered abnormal in one context may not be abnormal in another.

In summary, while both terms describe deviations from expected patterns in data, *anomaly* is more specific and often refers to unexpected patterns or outliers, whereas *abnormality* is a broader term that includes anomalies along with other types of deviations from the norm.

4.4.2 Holistic Pedestrian-agnostic Approach

A general way to characterize the behavior of pedestrians in a given scene is by taking just the motion within that scene into account. The fundamental idea is to analyze the OF representing the overall motion. This also means that by pursuing this approach, such methods can be seen as *pedestrian-agnostic* since they do not need to be informed about the position of a person and the fact that the regarded area of the current frame shows a person as long as they are moving. However, this also implies that any other motion like such generated by the surrounding, i.e., moving objects like vehicles, plants, flags and various others, is taken into account. In order to evaluate the motion in a given scene, an approach for frame-based anomaly detection in surveillance scenarios, which is based on the cross-channel approach proposed by Ravanbakhsh et al. [Rav17a], is presented in this section. The given reference approach is extended using cycle-consistency [Zhu17] and morphological [Ser86] operations to address certain of its drawbacks as presented initially in [Gol19c]. In this section, the two corresponding domains are the appearance domain $\mathcal{D}^A$ and the motion domain $\mathcal{D}^B$. For the sake of clarity, only one direction of the training cycle of the GAN setup is presented on the following pages, namely from $\mathcal{D}^A$ to $\mathcal{D}^B$. The training of the opposite direction, i.e., the transfer from motion domain to appearance domain, is performed analogously.

4.4.2.1 Cross-Channel Approach

In order to implement the desired transfer between appearance and motion, a GAN-based approach is presented, which takes single camera frames and

optical flow maps between two consecutive frames to learn a mapping between both of them. This cross-channel approach builds upon the pix2pix-architecture [Iso17], which consists of a conditional Generative Adversarial Network (cGAN) to transfer images from source to target domain. Since the pix2pix expects both domains to be from the image domain, the OF fields have to be converted into images first. Therefore, each OF vector is transformed into HSV color space, resulting in a colored image with 3-channels, with the orientation of the vector being represented as hue and the length as value. Saturation is set to 1.0. Just like the classical GAN, cGANs consist of a generator G and a discriminator D, where the former aims to capture the real data distribution and the discriminator's task is to distinguish between real and generated data samples. In the given situation, both, the conditional generator and the conditional discriminator, get the source image $\mathbf{x}_A$ as additional information. The generator aims to compute the corresponding sample $\mathbf{x}_B$ in the target domain as output using random noise $\mathbf{x}_Z \sim p(\boldsymbol{X}_Z)$. In particular, two variants are presented, which enforce this transfer being successful, one by using the loss $\mathcal{L}^{(\mathrm{V})}_{\mathrm{cGAN}}(G,D)$ as introduced for the VanillaGAN [Goo14] and adapted to the cGAN setup:

$$\begin{aligned}\mathcal{L}^{(\mathrm{V})}_{\mathrm{cGAN}}(G,D) = \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_B\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_B)}[\log D(\mathbf{x}_B \mid \mathbf{x}_A)] \\ + \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_Z\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_Z)}[1 - \log D(G(\mathbf{x}_Z \mid \mathbf{x}_A))]\end{aligned} \tag{4.19}$$

and the other one using the least squared error [Mao17] as GAN loss, which is referred to as LSGAN throughout this thesis. In this respect, the loss function is split into a loss for the generator $\mathcal{L}^{(\mathrm{LS})}_{\mathrm{cGAN}}(G)$ and a loss for the discriminator $\mathcal{L}^{(\mathrm{LS})}_{\mathrm{cGAN}}(D)$:

$$\mathcal{L}^{(\mathrm{LS})}_{\mathrm{cGAN}}(G) = \frac{1}{2} \cdot \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_Z\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_Z)}\big[\big(D\big(G(\mathbf{x}_Z \mid \mathbf{x}_A)\big) - 1\big)^2\big], \tag{4.20}$$

$$\begin{aligned}\mathcal{L}^{(\mathrm{LS})}_{\mathrm{cGAN}}(D) = \frac{1}{2} \cdot \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_B\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_B)}\big[\big(D(\mathbf{x}_B \mid \mathbf{x}_A) - 1\big)^2\big] \\ + \frac{1}{2} \cdot \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_Z\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_Z)}\big[\big(D\big(G(\mathbf{x}_Z \mid \mathbf{x}_A)\big)\big)^2\big]\end{aligned} \tag{4.21}$$

Additionally, the GAN task is extended by adding the L1-distance between the generated output and the image $\mathbf{x}_B$ in the target domain to the loss function with the aim of artifact reduction as well as low frequency correctness [Zhu17, Iso17]. This loss is calculated according to:

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{\{\mathbf{x}_A,\mathbf{x}_B,\mathbf{x}_Z\}\sim p(\boldsymbol{X}_A,\boldsymbol{X}_B,\boldsymbol{X}_Z)}\left[\| G(\mathbf{x}_Z \mid \mathbf{x}_A) - \mathbf{x}_B \|_1\right] \tag{4.22}$$

These losses are added up in a weighted sum as introduced in the original pix2pix GAN implementation [Iso17]. The whole setup is used to train generators, which are able to transfer normal camera scenes in expected optical flow maps and vice versa. The corresponding training of the GAN is performed for both directions in an alternating manner. In case of transferring appearance to motion information, the optical flow map $\tilde{\mathbf{x}}_{B,t}$ is generated based on the corresponding frame $\mathbf{x}_{A,t}$. The ground truth flow map $\mathbf{x}_{B,t}$ is computed based on the frame $\mathbf{x}_{A,t}$ and $\mathbf{x}_{A,t+1}$. During application, abnormal scenes, i.e., scenes that were not represented in the training data, are transferred poorly which leads to a higher disparity between the target image and the generated one. The aforementioned disparity is then used to calculate an anomaly score.

Figure 4.18: Schematic architecture of the VGG16 with highlighted utilized sub-architecture. The faded layers, starting with the first convolutional layer of the fourth block, were removed from the feature extractor, so that the output of the first layer of the fifth block of the architecture generates the features for the semantic comparison of frames.

The general workflow can be described as shown in the following. Firstly, for comparing optical flow maps the frame-wise anomaly score is calculated directly based on the disparity between the translated and the original flow map. Secondly, different to [Rav17a] who uses an AlexNet [Kri12] the first layer of the fifth convolutional block (5-1) of a pre-trained VGG16 network [Sim14] is used in order to extract a flattened feature vector from video frames in

order to determine a semantic difference between two frames. Finally, the disparity between the feature map of the original and the transferred image is used as anomaly score. In both cases the differences of the feature maps are calculated element-wise, squared and afterwards summed up along the channel dimension. The result of the corresponding operation is a heatmap $\Delta \in \mathbb{R}^{m \cdot n}$ of the squared differences where m and n denote the spatial dimensions. These heatmaps are utilized to calculate the frame wise anomaly score function $\alpha : \mathbb{R}^{m \cdot n} \rightarrow \mathbb{R}_0^+$ as follows:

$$\alpha(\Delta) = \sqrt{\frac{1}{m \cdot n} \sum_{i=0}^{m \cdot n - 1} \Delta^{(i)}} \tag{4.23}$$

where $\Delta^{(i)}$ denotes the ith entry of the heatmap Δ. Thus, the anomaly score is based on the root mean squared error between the original and the transferred representation in each domain. As in other work [Rav17a, Rav17b, Lee18, Liu18], anomaly scores are normalized video-wise. The workflow as described above yields two heatmaps for each sample, one for each of the two domains, therefore both heatmaps are fused as proposed in [Rav17a]. The fused heatmap $\bar{\Delta}_F$ is hence determined as shown below.

$$\bar{\Delta}_F \coloneqq \bar{\Delta}_C + \lambda_h \cdot \bar{\Delta}_O \tag{4.24}$$

It is the result of a weighted sum of the video-wise normalized heatmaps obtained from translation from camera frame to OF $\bar{\Delta}_O$ and from OF to camera frame $\bar{\Delta}_C$.

4.4.2.2 Cycle-Consistency Extension

Since the Cycle-GAN has proven its effectiveness for DA and other tasks, the adapted model is extended by cycle-consistency as proposed in [Zhu17] and introduced in Section 3.3.1. A given sample $\mathbf{x}_A$ taken from a source domain $\mathcal{D}^A$ is translated to a target domain $\mathcal{D}^B$ using the generator G_{AB}. The counterpart generator G_{BA} of the cross-channel approach aims to reconstruct the translated input sample in its original domain. The cycle-consistency loss $\mathcal{L}_{\text{cyc}}$

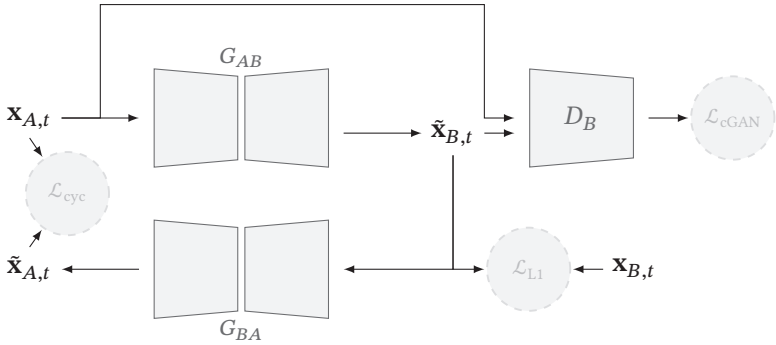


Figure 4.19: Schematic overview of the GAN training. During the training process of the GAN-approach two generators are trained. This schematic illustrates the direction from appearance to motion, where the generators G_{AB} and G_{BA} are trained one for each transformation direction between motion and appearance domain. The procedure from motion to appearance is identical.

is the pixel-wise L1 distance between the input and its reconstruction:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\mathbf{x}_A,\mathbf{x}_Z}\left[\| G_{BA}(\mathbf{x}_Z \mid G_{AB}(\mathbf{x}_Z \mid \mathbf{x}_A)) - \mathbf{x}_A \|_1\right] \tag{4.25}$$

where $\mathbf{x}_Z$ represents random noise, which is achieved by using dropout within the network as introduced by Isola et al. [Iso17]. The weighted cycle-consistency loss is added to the loss for the domain transfer as introduced in Section 4.4.2.1. Substituting the corresponding generators in the loss terms leads to the composite objectives for VanillaGAN and LSGAN as shown in Equations (4.26) to (4.28). The overall schematic is shown in Figure 4.19.

$$\mathcal{L}^{(V)} = \mathcal{L}^{(V)}_{\text{cGAN}}(G_{AB}, D_B) + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{L1}}(G_{AB}) + \lambda_{\text{cyc}} \cdot \mathcal{L}_{\text{cyc}} \tag{4.26}$$

$$\mathcal{L}^{(LS)}_{G} = \mathcal{L}^{(LS)}_{\text{cGAN}}(G_{AB}) + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{L1}}(G_{AB}) + \lambda_{\text{cyc}} \cdot \mathcal{L}_{\text{cyc}} \tag{4.27}$$

$$\mathcal{L}^{(LS)}_{D} = \mathcal{L}^{(LS)}_{\text{cGAN}}(D_B) \tag{4.28}$$

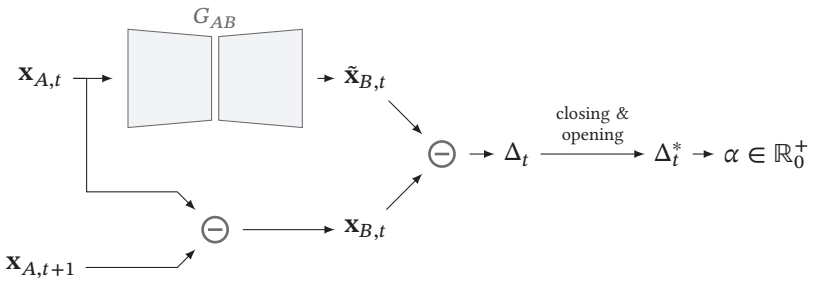


Figure 4.20: Illustration of the inference process. $\mathbf{x}_{A,t}, \mathbf{x}_{A,t+1} \in \mathbb{R}^{n \cdot m}$ are video frames at two consecutive timesteps based on which $\mathbf{x}_{B,t}, \tilde{\mathbf{x}}_{B,t} \in \mathbb{R}^{n \cdot m}$ are computed, i.e., the current and a predicted optical flow respectively. $\Delta_t \coloneqq (\mathbf{x}_{B,t} - \tilde{\mathbf{x}}_{B,t})^2$ captures the difference in both flows, which is then refined using morphological operations. Finally, the anomaly score is determined as $\alpha \coloneqq \sqrt{(n \cdot m)^{-1} \cdot \|\Delta_t^*\|_1}$.

4.4.2.3 Noise Suppression

So far the resulting heatmaps are taken as they are. However, certain structures or even noise can lead to minor potentially large activated spots in these heatmaps. In the field of application regarded within this thesis, anomalies mostly exhibit larger regions so that smaller areas can be ignored. To do so, a post-processing step follows the computation of the disparities, which gets rid of such *noise*. This is achieved by applying classical morphological operations, namely *closing* and *opening* [Ser86] on these heatmaps. Therefore, Δ has to be interpreted as a binary image first, which is achieved by clipping the values of each bin $\Delta^{(i)}$ between 0 and 1 using the following formula

$$f_{\text{clip}}(\Delta^{(i)}) \coloneqq \begin{cases} 1 & \text{if } \Delta^{(i)} > 0 \\ 0 & \text{if } \Delta^{(i)} = 0 \end{cases} \tag{4.29}$$

Closing is then applied to eliminate small holes from large area segments, which makes large area differences robust to opening which is applied afterwards to eliminate the small area differences and thus reduce false positive predictions. The resulting refined heatmap is denoted as Δ^*. For an overall overview of the workflow during inference see Figure 4.20.

4.4.3 Skeleton-based Behavior Analysis

Using a holistic approach as presented in the preceding section offers various benefits as it is general and can be applied to almost any kind of motion-based behavior or movements, like overall traffic patterns generated by the involved vehicles. However, at the same time such GAN-based methods suffer from various drawbacks, like view dependency, the challenging training and computational performance issues that easily result in inferences taking too long to being applied in a live setup. Furthermore, interpreting the output of a GAN and the resulting heatmaps can be rather difficult.

Especially the last two issues are addressed by applying a human-centric approach. Skeleton features are low-dimensional and human-comprehensible semantic features that focus solely on the body structure. They are easier to understand than differences on frames or OF, and are human-centric, which excludes other objects from the computation of an anomaly score for a given situation. Therefore, following the pedestrian-agnostic approach presented before, this section takes a look on another method to overcome privacy-concerns and still being somehow human-centered. Inspired by the achievements of graph-convolutional methods in the field of Skeleton-based Action Recognition (SBAR), as demonstrated in recent work by Yan et al. [Yan18], Yu et al. [Yu18], and Liu et al. [Liu20], this section examines the feasibility of applying these methods to the task of Skeleton-based Anomaly Detection (SBAD). Thus, this thesis presents a model that utilizes so-called Spatio-temporal Graph Autoencoder (ST-GCAE) modules and falls within the category of reconstruction-based techniques and operates at a microscopic scale.

Note that both, SBAR and SBAD are sub-fields of the Skeleton-based Behavior Analysis (SBA). Methods for SBAR focus on different classes of behaviors and actions, like *kicking*, *running*, *waving*, *walking* and many more. SBAD on the other hand tackles the categorization of behavior into *normal* and *abnormal* behavior. It is motivated mainly through the field of anomaly detection, which is most often an unsupervised learning approach, and hence can be seen as rather uninformed or partly informed. Of both fields the former is not

regarded within this thesis, whereas the latter matches the needs and desires of police authorities.

4.4.3.1 The Binary Autoencoder

As mentioned above, the approach works on temporal skeleton sequences obtained by an HPE algorithm. The way the Binary Autoencoder (BinAE) is designed is inspired by the work of researchers like Ionescu et al. [Ion19] and Chang et al. [Cha20], where multiple AEs are applied in parallel in order to extract features focusing on different aspects, like motion or appearance. For this work, the concept has been adapted for the use case of supervised training of a model for detecting anomalous human behavior. The model proposed is called BinAE as it consists of two branches, with each being an AE. Essentially, both AEs are build upon the spatio-temporal module presented by Markovitz et al. [Mar20], which is the primary component of the ST-GCAE. This module use two kind of graph-convolutions, one for spatial and the other for temporal information. The overall model architecture described in the following, with integrated memory modules, is visualized in Figure 4.21. One of the AEs is trained to reconstruct only normal human behavior, while the other one only learns to reconstruct abnormal human behavior. At inference time, the input sample $\mathbf{x} \in \mathbb{R}^{T \times K \times 3}$ will be passed to both AEs. The embeddings $\mathbf{h}_{\mathrm{n}} \in \mathbb{R}^{c}$ and $\mathbf{h}_{\mathrm{a}} \in \mathbb{R}^{c}$, where $c \in \mathbb{N}$ is the embedding dimension, calculated by the normal and abnormal AE, respectively, are concatenated and fed into an MLP which consists out of two fully-connected layers.

In order to distinguish both AEs, in the following normal and abnormal AE are referred to as AE_{n} and AE_{a} respectively. The first layer of the MLP receives the concatenated AE outputs and projects them to a dimension of 128, followed by a ReLU activation which is used between the fully-connected layers. The second fully-connected layer consists of one output neuron followed by a sigmoid activation which yields anomaly scores in the range of $[0, 1]$. The underlaying assumption is that the corresponding AE will fail to properly reconstruct those skeleton sequences not belonging to the respective class since those were not seen during training. This implies, that the AEs will not be able

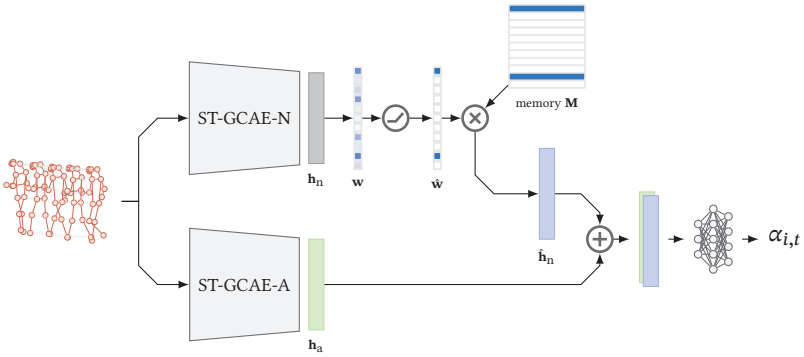


Figure 4.21: Illustration of the skeleton-based behavior analysis. Input sequences of fixed size are encoded as $\mathbf{h}_n$ and $\mathbf{h}_a$ by two identical encoder networks. Inspired by [Gon19] the branch for the normal behavior uses a learned content addressable memory [Wes15, Rae16] to obtain $\hat{\mathbf{h}}_n$ from $\mathbf{M} = (\mathbf{m}_i)_{i=0,\dots,L-1}^{T}$ instead of using the generated embedding $\mathbf{h}_n$ directly for better generalization.

to compress the right information to the embedding. Since the embedding holds the essential information to reconstruct the data, the subsequent MLP has access to the necessary information, which allows the network based on the embeddings to decide whether the input is an abnormal or normal sample.

4.4.3.2 Memory Extension

When trained using normal scenarios in the setting of video surveillance, the AE_n tends to reconstruct the input too well, even if the AE was only trained on normal samples, it is able to even reconstruct abnormal samples [Gon19]. Therefore, the embeddings of the two AEs are not different enough to be distinguished by the MLP. In order to control this, a memory module is added to regularize the process. The essential idea is, that by using a memory that was learned on normal samples, these cannot be used to reconstruct an anomalous sample, yielding higher loss values and hence improving the performance.

As defined by Gong et al. [Gon19] a memory module holds L vectors of size $c \in \mathbb{N}$ representing the most common normal patterns seen during model training. The memory is represented as a matrix $\mathbf{M} \in \mathbb{R}^{L \times c}$, which is placed

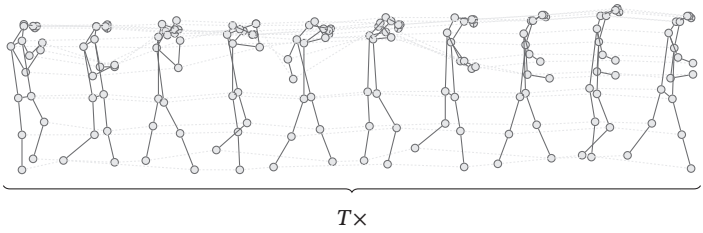


Figure 4.22: Visualization of input sample $\mathbf{x} \in \mathbb{R}^{T \times K \times 3}$. In the spatio-temporal input volume, $T \in \mathbb{N}$ corresponds to the number of regarded timesteps and $P \in \mathbb{N}$ to the number of keypoints in the chosen human body model. Solid edges are spatial connections defined by the body model, and dashed edges indicate a temporal connection between successive timesteps.

between the encoder and decoder module of an AE. Therefore it is easily possible to integrate such module into the ST-GCAE architecture and is hence applied after the encoder calculates the hidden representation $\mathbf{h} = f(\mathbf{x})$ of the corresponding input sample $\mathbf{x} \in \mathbb{R}^{T \times K \times 3}$.[1] For a better understanding, Figure 4.22 illustrates the spatio-temporal input $\mathbf{x}$.

Internally, the memory module uses the incoming embedding $\mathbf{h}$ as a query vector, with which in combination with an addressing vector $\mathbf{w} \in \mathbb{R}^L$ the network obtains the altered hidden vector $\hat{\mathbf{h}}$. This can be mathematically formulated as

$$\hat{\mathbf{h}} \coloneqq \mathbf{w} \cdot \mathbf{M} = \sum_{i=0}^{L-1} w^{(i)} \cdot \mathbf{m}^{(i)} \tag{4.30}$$

where $w^{(i)} \geq 0 \; \forall w^{(i)} \wedge \sum_{i=0}^{L-1} w^{(i)} = 1$. Here, $\mathbf{m}^{(i)}$ represents the ith memory item in $\mathbf{M}$ and $w^{(i)}$ are the items of the addressing vector.

[1] The values are set to $T = 12$, $K = 18$, and $L = 2{,}000$.

Contrary to what is stated in [Gon19], not cosine similarity but an inner product combined with softmax is used in their implementation in order to calculate the addressing vector $\mathbf{w}$, so it will be used here as well:

$$\mathbf{w} \coloneqq f_{\text{softmax}}(\mathbf{h} \cdot \mathbf{m}^{(i)}) \tag{4.31}$$

Furthermore, the authors propose to restrict the addressing vector by enforcing a sparse representation in order to avoid complex combinations being used to reconstruct abnormal samples. To do this, all entries $w^{(i)}$ of $\mathbf{w}$ below a certain threshold λ are clipped to 0, which is achieved by applying the following function:

$$\hat{w^{(i)}} \coloneqq \frac{\max(w^{(i)} - \lambda, 0) \cdot w^{(i)}}{|w^{(i)} - \lambda| + \epsilon} \tag{4.32}$$

Here, $\epsilon \in \mathbb{R}^+$ is a small positive number. Gong et al. [Gon19] state that applying this shrinkage operation leads to the model learning a representation made up of fewer, more relevant memory items and makes the memory $\mathbf{M}$ adapt to this sparse addressing. The authors advise to set the shrinking threshold λ for the sparse attention vector to $\left[\frac{1}{n}, \frac{3}{n}\right]$ [Gon19]. In the models used here, it is always set to $\frac{1}{n}$. The memory items are learned end-to-end during training together with the encoder and decoder and are kept fixed during testing.

4.4.3.3 Model Training

The training process for the BinAE is separated into two stages: During the pre-training stage only the two AEs are trained, where each of the AEs is optimized to reconstruct its corresponding samples, i.e., one for normal and the other for anomalous samples. The corresponding objective is given in Equation (4.33), which essentially consists of two parts, the reconstruction error $\mathcal{L}_{\text{rec}}$ between input $\mathbf{x} \in \mathbb{R}^{T \times K \times 3}$ and its reconstruction $\bar{\mathbf{x}} \in \mathbb{R}^{T \times K \times 3}$, and the entropy loss for the memory $\mathcal{L}_{\text{entropy}}$, which tries to create a sparse

addressing vector $\hat{\mathbf{w}}_k = (\hat{w}_k^{(0)}, \ldots, \hat{w}_k^{(L-1)})^\intercal \in \mathbb{R}^L$.

$$\mathcal{L}_{\text{MemAE}} := \frac{1}{N} \sum_{i=0}^{N-1} \Big\{ \underbrace{\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2}_{=: \mathcal{L}_{\text{rec}}} - \lambda \cdot \underbrace{\sum_{l=0}^{L-1} \hat{w}_i^{(l)} \cdot \log(\hat{w}_i^{(l)})}_{=: \mathcal{L}_{\text{entropy}}} \Big\} \tag{4.33}$$

with $\lambda := 2 \cdot 10^{-4}$ as proposed by Gong et al. [Gon19] and used in their presented Memory-augmented Autoencoder (MemAE), and N being the number of samples in the current training batch. While the AE_n is trained using the objective presented in Equation (4.33), the AE_a just uses the $\mathcal{L}_{\text{rec}}$ term as its loss function. For further details on the memory unit, please refer to Gong et al. [Gon19].

The second stage takes the pre-trained AEs and continues the training with the attached MLP based on the concatenated embeddings $\hat{\mathbf{h}}_\text{n}$ and $\mathbf{h}_\text{a}$. Since the goal of step 2 is to output an anomaly score $\alpha \in [0,1]$, where $\alpha = 1$ for abnormal samples and $\alpha = 0$ for normal samples, the loss function used for training is chosen as the Mean Squared Error (MSE), which is given in Equation (4.34).

$$\mathcal{L}_{\text{MSE}} := (\alpha - \bar{\alpha})^2 \tag{4.34}$$

where $\alpha \in [0,1] \subset \mathbb{R}$ is the output of the BinAE and $\bar{\alpha} \in \{0,1\}$ is the target value.

As such datasets are typically dominated by samples that are easy to classify since they are often overrepresented, e.g., walking pedestrians, such samples should have an lower influence on the training as soon as they are recognized by the model. It is modeled using in the focal loss [Lin18], which replaces the MSE as used in the initial model [Gol22b]. Lin et al. [Lin18] introduced the focal loss as

$$\mathcal{L}_{\text{focal}} := -\beta_t \cdot (p_t)^\gamma \cdot \log(p_t) \tag{4.35}$$

with

$$p_t := \begin{cases} p & \text{if } \bar{\alpha} = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{4.36}$$

where $\beta, \gamma \in \mathbb{R}$ are two parameters controlling the overall shape of the loss curve. For applying the focal loss to a regression problems Lu et al. [Lu20] adapted the notion of the original loss and presented the following formula

$$\mathcal{L}_{\text{focal}} := d^{\gamma} \cdot d^2 = d^{\gamma+2} \tag{4.37}$$

where $d := |\alpha - \bar{\alpha}|$. An extension to the focal loss is the so-called shrinkage loss [Lu20], which tackles the weighting issue of the naive interpretation of focal loss for regression application, namely that not only easy samples are affected but also hard samples, even if the re-weighting is less aggressive. The shrinkage loss is defined as follows:

$$\mathcal{L}_{\text{shrink}} := \frac{d^2}{1 + \exp(a \cdot (c - d))} \tag{4.38}$$

Both hyperparameters $a, c \in \mathbb{R}$ are kept as proposed by Lu et al. [Lu20], i.e., the values are set as $a := 10$ and $c := 0.2$.

This model can only be used when both normal and abnormal samples are available during training. Furthermore, the model expects class labels for the samples, so that the samples can be passed into the appropriate AE, resulting in an overall model that falls into the category of supervised trained models. To address this circumstance, an altered version of the BinAE is proposed, which varies in the way both AEs are combined. The so-called DualHeadAE merges both AEs with respect to the encoder modules and keeps both decoder modules. Figure 4.23 illustrates the resulting architecture.

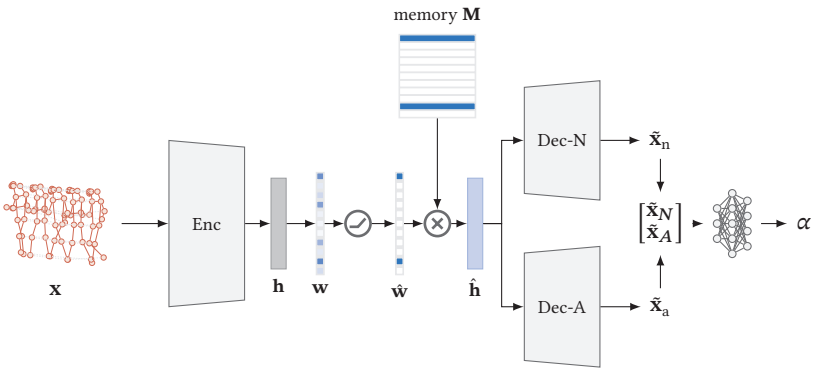


Figure 4.23: Alternative version of the BinAE that allows for full end-to-end learning. This model shares the same architecture for components of the AEs and the MLP with the original BinAE and is referred to as DualHeadAE. It comes with a minor architectural change, where the encoder is shared between normal and abnormal samples.

Since the DualHeadAE allows for end-to-end learning, the training objective is slightly altered combining the original losses in one. This is shown in Equation (4.39).

$$\mathcal{L} \coloneqq \bar{\alpha} \cdot \mathcal{L}_{\text{rec}}^{(a)} + (1 - \bar{\alpha}) \cdot \mathcal{L}_{\text{MemAE}} + \mathcal{L}_{\text{shrink}} \tag{4.39}$$

This way the DualHeadAE can be seen as an extension to the BinAE as presented initially in [Gol22b].

4.4.3.4 Pseudo-Anomalous Samples

So far, this section sheds light on the developed model and its architecture as well as on the general training process. In order to support the model during training and not solely relying on existing positive samples, i.e., anomalies, the training process is supported with automatically generated pseudo-anomalies. To do so, the existing normal samples are altered to generate such pseudo-anomalies that should be ideally identified by the network as anomalous samples. This approach is based on the idea of Wang et al. [Wan22] where

the authors proposed to divide videos into slices, which on the other hand are split into temporal and spatial patches. These patches are then used to generate two permutations of each input sample, namely a temporal reordered one, and a spatially reordered sample. This idea is incorporated in the DualHeadAE model, with two distinctions:

(i) the DualHeadAE is not trying to reconstruct the initial order of the passed sample, and

(ii) uses them on the other hand in order to learn and focus clearly on the underlaying representation of normal samples.

Since the DualHeadAE does not work with videos and images as inputs, adapting the idea from above demands for changes in order to generate such permutations. This process is visualized in Figure 4.24, where the two permutations $\pi_i : X \to X$ with $X := \mathbb{R}^{T \times K \times 3}$ and $i \in \{\text{spat}, \text{temp}\}$ are exemplarily given and illustrated. The idea for temporal reorder can be applied almost directly to sequences of poses. Each pose $\mathbf{V}_t$ at timestep $t \in \{0, \dots, T-1\}$ is firstly transformed to be center-free, shuffled and the new poses for timestep t is transformed back to original position of the old pose at timestep t. This temporal permutation is then referred to as $\pi_{\text{temp}}(\mathbf{V}_t)$. Not rearranging the single poses spatially could simplify the task for the network to just look at the continuity of the overall motion of a pose.

However, the spatial permutation $\pi_{\text{spat}}(\mathbf{x})$ is generated differently for pose sequences, which is visualized in Figure 4.25. Therefore, each limb of the human body is interpreted as a "patch", meaning that there are four patches per person. The head-related keypoints, as well as the torso-related ones are kept unaltered, since especially the torso keypoints are taken as reference to assemble pseudo-abnormal samples. As shown in Figure 4.25, each limb is separated from the pose and re-located to another random position, where it is attached to the new anchor. Figure 4.24 shows this for a full sequence, where for every timestep the same re-order of limbs is performed, while preserving the temporal order. This is done for the whole sequence, i.e., for each timestep in the same way meaning that $\mathbf{x}$ and $\pi_{\text{spat}}(\mathbf{x})$ differ solely with respect to the configuration of body parts.

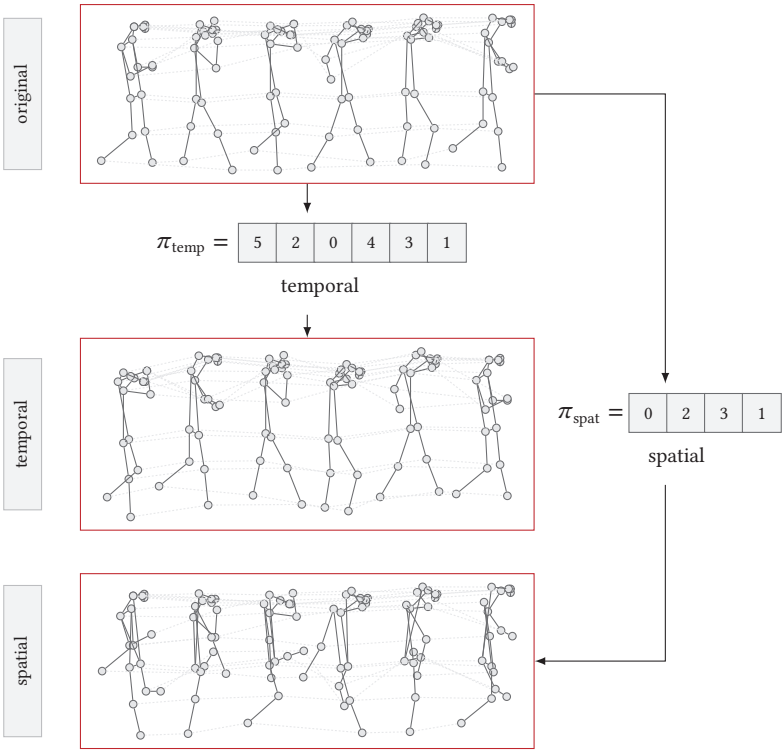


Figure 4.24: Given a pose sequence $\mathbf{x} \in \mathbb{R}^{T \times K \times 3}$ two kinds of alterations are performed independently, namely random reorder π_{temp} of the single timesteps, and random reorder of body parts π_{spat}, in particular an exchange of different limbs. These two cases are shown above.

4.4.4 Summary

This section gave in introduction into the field of behavioral analysis, starting with a definition of the problem space. The subsequent sections presented different approaches to analyze human behavior from video data, one following an own holistic approach based on a GAN setup and the other one tackling the problem from a human-centric perspective using skeletal representations of humans. Especially for the the latter different extensions were introduced,

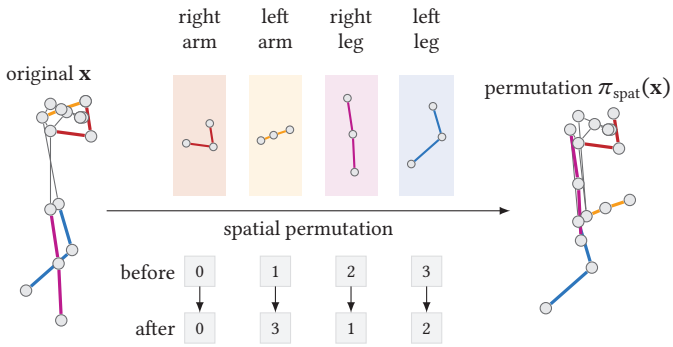


Figure 4.25: Visual explanation of the spatial pose-permutation π_{spat}. Subject to the alteration are the limbs of the body model, whereas the torso and head are kept as they are. The body parts are separated from the torso at their respective anchor points (i.e., shoulders and hips), shuffled and re-attached at a (possibly) different anchor point. The old and new position for the given example are illustrated in the lower part of the figure.

in particular two ways of using synthetic data, either the MixAMoR data or by generating permutations. Both address the aspect of having appropriate data that can be used to train a model in a privacy-friendly way. These methods and model variants will be evaluated and compared in the next chapter, where the holistic will serve as a baseline as it is less versatile compared to the skeleton-based approach.

5 Experiments

5.1 Overview

This chapter presents the assessment procedures that define the blueprints for the conducted experiments including their results. To capture the content of this dissertation, the chapter is divided into three parts. Beginning with Section 5.2, which presents various evaluation metrics that are used to measure the performance of the different approaches, the chapter continues by presenting the systematic evaluation protocols for the various tasks addressed in this thesis is presented in Section 5.3, outlining the applied methodology. The initial focus in Section 5.2 lies on introducing key metrics employed for performance evaluation. Subsequently, a systematic evaluation protocol for the various tasks addressed in this thesis is presented in Section 5.3, outlining the applied methodology. Finally, Section 5.4 presents the resulting findings including a concise and thorough discussion.

5.2 Metrics for Performance Assessment

5.2.1 Human Pose Estimation

In order to assess the performance of an HPE method, a quantitative measure for the performance is needed that evaluated the algorithm's capabilities.

As for any other detection tasks, the resulting outputs of the keypoint detector require a categorization of each into different categories. This is broadly

known as a *confusion matrix*, a special kind of contingency matrix. The mentioned categories are presented in the following list.

- *True Positive (TP)*: correctly detected keypoint
- *True Negative (TN)*: correctly not detected keypoint
- *False Positive (FP)*: unmatched detected keypoint
- *False Negative (FN)*: unmatched ground-truth keypoint

As for other detection tasks, the prevalent kinds of measures for human pose estimation are *precision* and *recall*. Precision measures the fraction of correctly predicted keypoints whereas recall measures the fraction of ground truth keypoints that were detected.

Let V be a set of keypoints and $f_{\text{split}} : V \to \mathcal{P}(V)^4$ be a function that divides V into four different subsets representing the mentioned categories. The resulting split[1] $\mathcal{V}$, meets the following conditions

$$\mathcal{V} \coloneqq \{\, V_k \mid V_k \in \mathcal{P}(V),\ \forall k \in \{\text{TP}, \text{TN}, \text{FP}, \text{FN}\} \,\} \tag{5.1}$$

$$V_j \cap V_k = \varnothing \quad \text{with} \quad \forall j,k \in \{\text{TP}, \text{TN}, \text{FP}, \text{FN}\} \text{ and } j \neq k \tag{5.2}$$

$$V = V_{\text{TP}} \cup V_{\text{TN}} \cup V_{\text{FP}} \cup V_{\text{FN}} \tag{5.3}$$

Using these recall f_{rec} and precision f_{pre} are defined as

$$f_{\text{pre}}(\mathcal{V}) = \frac{|V_{\text{TP}}|}{|V_{\text{TP}}| + |V_{\text{FP}}|} \tag{5.4}$$

$$f_{\text{rec}}(\mathcal{V}) = \frac{|V_{\text{TP}}|}{|V_{\text{TP}}| + |V_{\text{FN}}|} \tag{5.5}$$

For calculating *precision* and *recall*, analogous to the IoU for object detection, a measure is needed to decide whether a pose is located correctly or not. The most commonly used similarity metrics are the Object Keypoint Similarity and the Probability of Correct Keypoint (head) [Goo16, Kal19]. Both are shortly introduced in the following.

[1] Note that $\mathcal{V}$ has not to be a partition of V, since the subsets $V_k \subset V$ can be empty.

5.2.1.1 Object Keypoint Similarity

The OKS was introduced by Ronchi et al. [Ron17] as a challenge metric for the COCO keypoint challenge. As the name OKS implies, the goal of this metric is to measure the similarity between two sets of keypoints, in general the annotated ground truth and the predicted keypoints of a keypoint detector. Given two sets of keypoints V_1 and V_2 with

$$V_n = \left\{ \mathbf{v}_{i,n} = \left(v_{i,n}^{(0)}, v_{i,n}^{(1)}, v_{i,n}^{(2)}\right) \mid \mathbf{v}_{i,n} \in \mathbb{R}^2 \times \{0,1,2\} \right\} \tag{5.6}$$

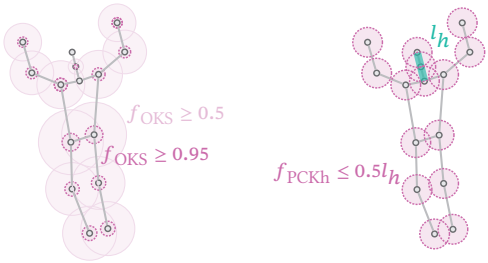
where $i \in \{0, \dots, N-1\}$ denotes the index of the keypoint. The OKS is then defined as

$$f_{\text{OKS}}(V_p, V_q) \coloneqq \frac{\sum_{i=0}^{N-1} \left[\exp(\frac{-d_i^2}{2s^2\kappa_i^2}) \cdot \mathbf{1}(\mathbf{v}_{i,q}) \right]}{\sum_{i=0}^{N-1} \left[\mathbf{1}(\mathbf{v}_{i,q}) \right]}, \tag{5.7}$$

with $\mathbf{1} : V \to \{0,1\}$ being the indicator function with

$$\mathbf{1}(\mathbf{v}) \coloneqq \begin{cases} 1 & \text{if } v^{(2)} > 0 \\ 0 & \text{if } v^{(2)} = 0 \end{cases} \tag{5.8}$$

where $d_i = \| \mathbf{v}_{i,p} - \mathbf{v}_{i,q} \|_2$ is the euclidean distance between two keypoints and $v_{i,q}^{(2)}$ the visibility of the ground truth keypoint. As Equation (5.7) shows, the similarity of a keypoint is determined by passing the euclidean distance of the keypoints through an unnormalized Gaussian distribution with standard deviation $s\kappa_i$, with s being the person scale and κ_i a keypoint constant. A perfect prediction will yield $f_{\text{OKS}}(V_p, V_q) = 1$ whereas predictions that are off by a few standard deviations $s\kappa_i$ will have $f_{\text{OKS}}(V_p, V_q) = 0$. The keypoint constant κ_i was determined empirically for each type of keypoint. Therefore, Ronchi et al. [Ron17] redundantly annotated 5,000 images and measured the per-keypoint standard deviation σ_i with respect to the persons scale. The keypoint constant differs largely depending on the type of keypoint, resulting in rather low deviation for such related to the facial region like eyes, ears, and nose, and larger spread for keypoints on knees or hips.



(a) OKS **(b)** PCKh

Figure 5.1: The right side shows the conceptual idea of the PCKh where the green line indicates the length of the head segment, i.e., the reference value for the matching areas, which are indicated by a dashed purple circle. The matching threshold l_h is identical for all keypoints. In contrast, OKS uses fixed and empirically determined values κ_i per keypoint. Note that the pose model shown here is the SyMPose model, but the original OKS was defined on the COCO model. This is the reason that there are no thresholds for two of the head keypoints.

In order to determine that a pose is located correctly, the OKS score needs to exceed a pre-defined keypoint threshold as illustrated in Figure 5.1a. The OKS serves as a similarity measure in keypoint detection in the same vein as the IoU does for the object detection [Ron17].

5.2.1.2 Probability of Correct Keypoint (head)

Based on the so-called Probability of Correct Keypoint (PCK) [Yan13], Andriluka et al. [And14] introduced the PCKh which extends the PCK. The PCK defines a keypoint as correctly located if the keypoint is within the range of $\alpha \cdot \max(h, w)$ with w and h being the width and height of the person bounding box and α being a relative threshold. A problem which arises using this metric is that it depends on the articulation of a pose due to the dependency of the size of the bounding box. PCKh alleviates this problem by defining the matching threshold as 50% of the head segment length $l_h \in \mathbb{R}^+$.

$$f_{\text{PCKh}}(\mathbf{v}, \tilde{\mathbf{v}}) := \| \mathbf{v} - \tilde{\mathbf{v}} \|_2 \tag{5.9}$$

5.2.1.3 Summary

Figure 5.1 illustrates side by side the differences of the keypoint matching areas for both, OKS (Figure 5.1a) and PCKh (Figure 5.1b). While PCKh determines matching keypoints using an identical threshold for all keypoints, OKS uses different thresholds for different parts of the body. These were determined empirically by Ronchi et al. [Ron17] for the COCO model, which has been presented in Figure 2.5.

Both introduced metrics, the PCKh and the OKS are used in the same manner as to calculate the *average precision (AP)*, *mean average precision (mAP)*, *average recall (AR)*, and *mean average recall (mAR)*. AP and AR are computed for different thresholds ranging between 0.0 and 1.0, whereas the mAP and mAR, as the name already indicates, are the mean values over the different AP and AR scores respectively. However, note that PCKh is actually a distance measure between keypoints, whereas OKS measures the similarity of two poses, which effects the way matches are determined.

Although, PCKh and OKS have been both widely used to model the IoU for keypoint detection, this thesis however provides results using the current state-of-the-art, which is the OKS. The PCKh was presented for the sake of completeness.

5.2.2 Behavioral Anomaly Detection

In order to assess the performance of such behavioral analysis suited measures are necessary. There exists a broad variety of different metrics as presented by Sharif et al. [Sha22], of which this section introduces the most widely used ones. The subsequent evaluation will adhere to this decision. Three primary metrics are employed to assess the performance of anomaly detection models [Paz22]: the Area under the Receiver Operator Characteristic curve (AUC-ROC), the Area under the Precision-Recall curve (AUC-PR), and the Equal Error Rate (EER). While none of these metrics singularly captures the complete picture of overall performance, each comes with its own set of strengths and

weaknesses. Together, these metrics contribute to a more comprehensive understanding of the algorithm's true performance. The mentioned categories are presented in the following list.

- *TP*: correctly classified anomaly
- *TN*: correctly classified normality
- *FP*: incorrectly classified normality
- *FN*: incorrectly classified anomaly

In the following all aforementioned metrics and measurements will be defined and explained further.

5.2.2.1 Area Under the ROC Curve

The AUC-ROC quantifies a binary classification model's performance by measuring the area under the *True Positive Rate (TPR)* versus *False Positive Rate (FPR)* curve at various thresholds. Thereby, the TPR and FPR are defined as

$$f_{\text{TPR}} \coloneqq \frac{TP}{TP + FN} \tag{5.10}$$

and

$$f_{\text{FPR}} \coloneqq \frac{FP}{FP + TN} \tag{5.11}$$

respectively. Higher AUC-ROC values suggest superior class separation. Although the Receiver Operating Characteristic (ROC) curve illustrates the trade-off between TPR and FPR [Fer18], AUC-ROC does not provide insight into the model's final decisions. It yields a single value and struggles to convey meaningful information about False Negative Rate (FNR), crucial for real-world applications. FNR, especially when an anomaly is wrongly classified as normal, is challenging to assess based solely on AUC-ROC. AUC-ROC's sensitivity to imbalanced data [He13], prevalent in anomaly datasets, further limits its optimal use, particularly when one class is overrepresented.

5.2.2.2 Area Under the Precision-Recall Curve

Precision measures the ratio of accurate positive predictions to all positive predictions, while *Recall* calculates the ratio of accurate positive predictions to all positive samples. They are formally defined as

$$f_{\text{prec}} \coloneqq \frac{TP}{TP + FP} \tag{5.12}$$

and

$$f_{\text{rec}} \coloneqq \frac{TP}{P} \tag{5.13}$$

The *Precision-Recall Curve (PR)* illustrates the trade-off between Precision and Recall, by the area under the curve (AUC-PR), summarizing the curve's information. AUC-PR is more effective than AUC-ROC in evaluating a model's prediction ability, particularly in highly imbalanced data [Sai15], as it considers the FNR (i.e., when the model classifies an anomaly as normal). This makes it valuable for assessing the minority class [He13], often representing anomalous behaviors in anomaly detection. Despite its strengths, AUC-PR, like AUC-ROC, is a single number that lacks direct insights into correct negative classifications and does not quantify the number of incorrect decisions made by a model. Therefore, AUC-PR, while beneficial, provides an incomplete understanding of a model's overall performance.

5.2.2.3 Equal Error Rate

The EER is a valuable metric, representing the point on the ROC curve where the FPR is equal to $(1 - f_{\text{TPR}})$ [Zha18a], indicating a better performance the lower its value is. By plotting FNR and FPR across various thresholds, intersecting curves reveal the EER, highlighting the threshold for a balanced trade-off between FNR and FPR. In video anomaly detection, the EER quantifies false

Table 5.1: Components of CrowDPB alongside with their categorization and purpose. Black circles indicate that the subset is used for the corresponding purpose. HPE and DA correspond to the experiments in Section 5.3.2, and VAD as well as SBAD to those in Section 5.3.3. Note that HGH18 has two parts, namely HGH18-HPE and HGH18-DA. They will be both referred to as HGH18 as long as there is no risk of mixing them up.

Subset	HPE	DA	VAD	SBAD
CrowdPE	●	○	○	○
SyMPose	●	●	○	○
HGH18	●	●	●	●
CaWa18	●	○	●	●
HeR19	○	○	●	●

alarms and missed anomalous frames at equilibrium. While EER alone provides limited insights into the overall performance of a model [Sul18], its integration with AUC-ROC and AUC-PR enhances overall model performance understanding.

5.3 Evaluation Protocol

In order to evaluate the impact and effects of the methodical considerations introduced in Chapter 4, this section presents the procedures to evaluate the presented approaches. First a custom dataset used for the experiments is introduced in Section 5.3.1, followed by two sections each focusing on one part of the presented methods. Beginning with the topic of HPE, Section 5.3.2 tackles the aspect of generating training data and presents the workflow of training an arbitrary human pose estimator using synthetically generated data and hence is related to Sections 4.2 and 4.3. This is followed by the behavioral analysis part as presented Section 4.4, which is subject to Section 5.3.3. Both parts include information on the training procedure, as well as facts on the evaluations, in particular a presentation of the employed metrics and datasets. Note that the reference procedures followed here were developed in collaboration with students and therefore have been introduced in [Bla19] and [Hoh24], for the task of HPE and SBAD respectively.

Table 5.2: Statistic for pedestrian and keypoint annotations in three selected subsets of CrowDPB. Due to the crowded character of all these datasets, being especially prominent for CaWa18 and HGH18-HPE, a significant number of annotated bounding boxes do not contain pose annotations.

Dataset	Frames	Bboxes	Keypoints		Tracks
			visible	occluded	
CaWa18	306	47,207	171,842	24,295	198
HGH18-HPE	800	72,299	57,131	13,793	189
CrowdPE	53	2,700	23,675	9,009	-
total	1,159	122,206	252,648	47,097	387

5.3.1 Crowded Dataset for Human Pose Estimation and Behavior Analysis

As mentioned in the introduction of this section, the first part presents the dataset used throughout the experiments. The Crowded Dataset for Human Pose Estimation and Behavior Analysis (CrowDPB) serves as quantitative and qualitative dataset to assess the performance on the different tasks in this thesis. As shown in Table 5.1 CrowDPB particularly consists of five subsets, each used for various (sub-)tasks. Starting with 25 frames and 833 bounding boxes in 2018 [Dis18, Gol19a], CrowDPB has been consequently extended not only by the overall number of images and human keypoint annotations, but also by increasing the overall number of pedestrians per frame. Initially used in [Gol19a] – presented as *CrowdPose*[1] at that time – the dataset was extended and used in further work [Kal19, Gol19b] under the name CrowdPE. Table 5.2 highlights various properties of those subsets that are used for the task of HPE. With respect to the available keypoint annotations, Table 5.3 shows the overall sGCI distributions as well as the average number of people per frame within each subset. In comparison to CrowdPE, both, HGH18-HPE and CaWa18, show a significant increase in crowdedness and hence emphasize the suitability of CrowDPB for evaluations on crowd-level. Figure 5.2 underlines this statement by showing the distribution of the sGCI for each subset.

[1] Was renamed due to the release of the public dataset CrowdPose by Li et al. [Li19].

Table 5.3: Analogous to Table 4.2, which compared the various sequences of SyMPose, this table shows the main characteristics of the three HPE-related subsets of CrowDPB. Although, all of the subsets show crowded scenarios, CaWa18 yields especially high crowdedness scores. This is mainly due to the kind of festivity.

	μ_{sGCI}	μ_{ppf}	σ_{ppf}	r_{ppf}
CrowdPE	0.61	50	32.54	5 - 130
HGH18-HPE	0.82	95	14.28	75 - 126
CaWa18	3.40	154	7.30	85 - 162

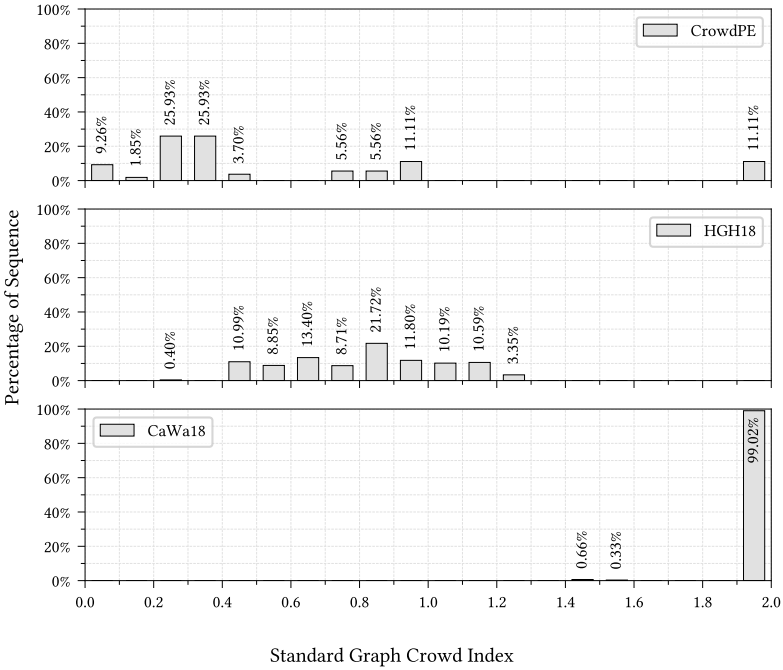


Figure 5.2: Distributions of the sGCI for three used datasets CrowdPE, HGH18-HPE, and CaWa18. CaWa18 shows that with respect to the sGCI it is the hardest dataset. Note: For the purpose of visualization, all values have been clipped between 0 and 2, which explains why CaWa18 has almost all frames falling into the last bin of the histogram. Table 5.3 indicates, that most sGCI values of CaWa18 are way higher.

5.3.2 Synthetic Data-driven Human Pose Estimation

This section focuses on the evaluation of the first methodical part with the aim of generating training data for the HPE task. As mentioned above, the overall procedure follows the considerations that were developed and presented in [Bla19, Gol20]. Firstly, Section 5.3.2.1 outlines preliminary considerations on the evaluation with respect to the datasets presented earlier, and Section 5.3.2.2 presents the exact configurations of the training. Finally, Section 5.3.2.3 then explains the assessment of the performance of the developed approach.

5.3.2.1 Datasets

As explained in Section 4.3.1 there are two domains that have to be considered. For the scope of this thesis, these domains are represented through two different datasets: The synthetic source domain is represented by the SyMPose dataset that has been introduced and described in Section 4.2.2, whereas the real-world target domain is represented by the HGH18-DA dataset that has been presented in Section 4.3.5. Although the datasets have been tailored for the exact purpose of the DA task, certain adjustments had to be made in order to stabilize training and facilitate the overall DA task.

Source domain. First and foremost, the SyMPose dataset contains a great variety of scenes. For instance, the illumination conditions vary within a wide range as the scenes were recorded at different in-game times of day and were generated containing various weather conditions. These conditions increase the tasks difficulty immensely, which is why only a subset of SyMPose is considered for the DA task. In particular, this subset consists of sequences S_1, S_3, S_7, S_9, S_{12}, S_{14}, and S_{23}, providing an overall number of 6,300 images in daylight situation, all located in urban setups. Figure 5.3 depicts characteristic frames for each of these sequences.

Target domain. Similar thoughts were given to the target dataset as well. By design, the set of training samples provided by the HGH18-DA shows far

Figure 5.3: Illustration of the overall appearance of the selected subset of SyMPose. It is obvious that the selection of sequences is characterized by an overall grayish impression which is beneficial to the adaptation process. The sequences are displayed in ascending order from left to right.

less variations in appearance compared to the dataset for the synthetic domain, but still there are parts of the dataset that deviate strongly from the rest. Since all cameras show very similar properties with respect to view and content it all comes down to the illumination of the scenes. In particular, those frames captured close to and after sunset show the strongest deviance from the rest of the dataset, which naturally comes from the changes in illumination. These include global lighting due to the absence of the sun, as well as local lights by the surrounding booths, both having a direct impact on the recorded video material in terms of contrast, and image noise. Since these kind of samples increase the difficulty of transferring between both domains for the developed model, frames recorded after 6 p.m. are not taken into account. Some random examples that show such difficult lighting situations are illustrated in Figure 5.4. These are characteristic examples, where the change in illumination creates various demanding and challenging situations that increase the difficulty for the generator to correctly adapt the appearance to the target domain. Moreover, the remaining samples of the HGH18-DA are randomly sub-sampled aiming for a final set of 3,000 samples, ensuring that there is still enough variety between samples. This procedure for the target domain conforms to the procedure for the WorldExpo '10 dataset followed in [Gol20].

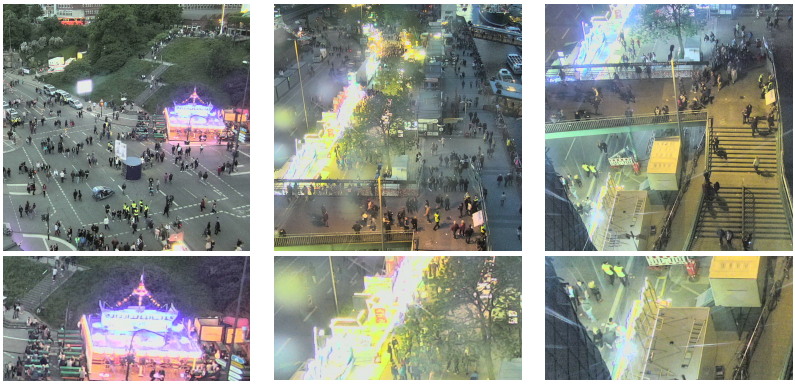


Figure 5.4: Examples of low-light frames from HGH18-DA. The upper row illustrates an overview over selected scenes, whereas the lower row shows particular close-ups highlighting certain aspects. Apparently, image contrast and clarity are directly affected by the lights. This can already be seen in the first image, where the overall illumination looks still homogeneous, but taking a closer look shows that the internal light compensation leads to a reduced dynamic range resulting in a loss of structural information as can be seen above the illuminated booth. The case illustrated in the middle shows, that the change of direction the light comes from creates visual flares as artifacts. These come in particular from combination with rain drops on the protective hood of the installed camera. As the third example shows, depending on the intensity of light and its angle of incidence into the protective hood, the hood itself can yield such visual artifacts.

Performance assessment. The last dataset that is relevant to the experiments for DA is the actual evaluation dataset. CrowDPB, as presented in Section 5.3.1, provides certain subsets that are suited for this kind of evaluation. The experiments are carried out using all of these HPE-related subsets, namely HGH18-HPE, CaWa18, and CrowdPE, if not stated differently, reporting results categorized for different crowdedness levels. This procedure is inspired by the initial approach proposed by [Li19] and adopted in [Bla19, Gol20].

5.3.2.2 Preliminary Considerations on the Training

The previous section explained what data is used as basis for the training procedure, and performance assessment. For the training of the domain adaptation frameworks, the images are augmented as depicted in Section 4.3.4,

where the fixed crop size is 512×512. All models are trained with a batch size of one and optimization is performed using the *Adam* optimizer [Kin14] with a fixed learning rate of 10^{-4} and $\beta_1 = 0.5$, $\beta_2 = 0.9$, which is a common optimizer configuration for GAN-training. For further technical details on the chosen optimizer see [Kin14]. The data augmentation techniques that were presented in Section 4.3.4 are utilized for all the DA experiments. Extensive hyperparameter tuning was discarded for reasons of lacking time and computational resources.

5.3.2.3 Evaluation Pipeline

In order to assess the quantitative performance of the DA approach, Blattmann [Bla19] introduced a task-specific evaluation protocol. Therefore, a HPE network is trained on various datasets, i.e., either synthetic, adapted or real-world data and the results on a common test set are taken as measure for the suitability of these images as training data. Different to Blattmann [Bla19], who used the SimpleBaselines [Xia18] model, this thesis employs the HRNet [Sun19] as pose estimation network. As this human pose detector solely serves as objective measure for the suitability of the adapted data as training examples, no hyperparameter tuning was applied to that network. For all of the datasets that are used as training sets, the model is trained for 40 epochs, with a batch size of 32 and an initial learning rate of 0.001, which is reduced by the factor of 10 at epochs 10, 20, and 35. To diminish the influence of stochastic effects, such as data shuffling, all pseudo random number generators are initialized with the same seed value, for the training procedure. Note also that only the HPE network is trained on the distinct training datasets, using ground truth bounding box detections. The same applies to the testing procedure.

In general, the evaluation can be done with an arbitrary pose estimator, e.g., OpenPose, AlphaPose, or ViTPose, as presented in Section 3.1.1. However, for convenience and consistency throughout this thesis HRNet was chosen, in particular version HRNetW48. The overall steps of the pipeline can be seen as follows:

(Step 1) Train DA model for 50,000 steps

(Step 2) Convert complete SyMPose-Urban dataset to target do main

(Step 3) Take the dataset and train HRNetW48 for 40 epochs

(Step 4) Use the best performing model and evaluate on evaluation dataset

Following this procedure results in identical-sized training datasets for the synthetic and adapted experiments. These datasets are the basis of the experiments.

5.3.3 Behavior Analysis

In literature there exists no unified way of evaluating SBAD approaches, which means that different research provides potentially different results for various datasets. This is particularly an issue with new datasets published that come up with specific ways to evaluate methods on. Therefore, this section presents a unified evaluation protocol, which is used to compare the approach presented in Section 4.4.3 with current state-of-the-art methods on different dataset. These datasets will be presented in their original form in Section 5.3.3.1. Since they either do not provide any evaluation procedure for SBAD yet or come with varying protocols, the actual evaluation procedure applied in this thesis is presented in Section 5.3.3.2. Note that this section mainly focuses on the evaluation of the skeleton-based approach, since the holistic approach (cf. Section 4.4.2) can be easily evaluated alongside with the former and does not need any additional data preparation.

5.3.3.1 Datasets

To measure the performance of skeleton-based methods different experiments are performed throughout this thesis. The main challenge with the Skeleton-based Behavior Analysis is the lack of suited datasets, which comes from the

extensive effort needed in order to annotate pose data, especially for video-based scenarios. This section presents various publicly available and commonly used datasets to evaluate the performance as well as three custom datasets that are part of the CrowDPB dataset introduced in Section 5.3.1. While the public datasets provide behavioral annotations, they are used for quantitative evaluation and the custom datasets are used for qualitative comparison.

5.3.3.1.1 ShanghaiTech Campus

The SHTC [Liu18] dataset was introduced by the *ShanghaiTech University* in June 2018. It was introduced as a video anomaly detection dataset, with the primary objective to enhance scene diversity by incorporating various camera angles and positions, while also capturing anomalies induced by abrupt motions like chases and brawls. In doing so, it distinguishes itself from previously released datasets, including the Avenue [Lu13] dataset, UCSD Pedestrian [Li13] dataset more precisely its two splits, *Ped1* and *Ped2*, and lastly the Subway [Ada08] dataset, showcasing its superiority in addressing these specific challenges.

The dataset consist out of 317,398 frames in total, where 274,515 and 42,883 frames are used for training and testing, respectively. With a video resolution of 856×480 at 24 frames per second it belongs to those datasets with rather small frame sizes. Anomalous activities happen in 17,090 frames, with 130 types of anomalous activities existing in a total of 13 scenes recorded with 13 different stationary mounted camera positions, all of which show complex light conditions and camera angles. These cameras recorded real-world scenarios on the campus of the university. In terms of annotations, the dataset provides frame-level and pixel-level anomaly annotations. As the dataset contains anomalous situations that are not associated with humans, there is a publicly available partition of the dataset, the so-called Human-Related ShanghaiTech Campus (SHTC-HR) dataset by Morais et al. [Mor19] that deals exclusively with anomalous activities associated with humans.

5.3.3.1.2 Vision-based Fallen Person

The Vision-based Fallen Person (VFP290K) [An21] dataset was published in October 2021 by the *Sungkyunkwan University* in South Korea. It addresses the challenging case of detecting fallen persons due to, various reasons, like health problems, violence or accidents. In total, it contains 294,713 frames of fallen persons in total, which were captured using a dynamically installed GoPro HERO5 camera, recorded in a resolution of 1080p at 25 frames per second. These frames are from 178 videos in total, including 131 scenes in 49 different locations. Therefore it provides a large-scale of collected fallen persons images in various real-world scenarios.

In addition, the dataset consists of a number of categorized subsets that organize the dataset and simplify working with it. There are two light conditions subsets, M_{day} and M_{night}, two camera heights subsets, M_{low} and M_{high}, which cluster the data in views in approximately 1 meter and 3 meters height, three different categories of various backgrounds M_{street}, M_{park} and M_{building}, a subcategory for the 49 locations where the videos were captured, M_{location}, and a subcategory of each location displaying the various camera viewpoints at these locations M_{scene}. There are a total of 15 actors, who are volunteering students, of whom three are international, switching clothes between scenes, to perform different fall scenarios. In order to ensure the accuracy of the annotations within the dataset, two strict annotation rules have been introduced to handle scenarios involving occlusion and overlap. The first rule, known as the occlusion rule, defines situations where a target's body is partially occluded by other objects or people. In cases where the body of a person is partially occluded but body parts of this persons are still visible, a bounding box is assigned that contains the these visible body parts of the person. The second rule, the overlap labeling rule, comes into play when one person occludes another, resulting in overlapping bounding boxes. In these instances, a bounding box is assigned to the front person only if the back person is occluded by more than 80%. To ensure compliance with these rules, the dataset maintains data quality assurance by cross-checking every single frame for each video by a team of five students.

Through the annotated bounding boxes the dataset provides pixel-level anomaly annotations, by labeling the bounding box of a fallen person as anomalous and non-fallen person as normal. For training and testing, the benchmark split introduced in [Gol22b] is adopted. Notably, none of the approaches analyzed in this work utilize a validation set, because of that the validation set of the aforementioned benchmark split is incorporated into the test set, such that the original test set and the validation set resemble the new test set.

In essence, the VFP290K dataset comes with a new challenge, namely the detection of falling persons in a real-world scenario. Furthermore, it is the only dataset where the frames are not captured from a stationary mounted camera, rather from a dynamic video recording and thus brings up additional challenges.

5.3.3.1.3 Charlotte Anomaly

The Charlotte Anomaly Dataset (CHAD) [Dan23] was initially published in December 2022 by the *University of North Carolina at Charlotte* [Paz22]. It is a high-resolution, multi-camera anomaly dataset in a commercial parking lot setting and the first dataset, which includes human bounding boxes, person identifier labels with person re-identification and human pose annotations for each actor. The multi-camera setting involves four stationary mounted cameras, where the videos of cameras 1-3 were recorded in 1080p at 30 frames per second and those of camera 4 were recorded in 720p also at 30 frames per second.

The CHAD dataset contains 412 videos with 1.15 million frames overall with over 87% normal and around 125,000 anomalous frames. To allow both supervised and unsupervised learning, the dataset includes two training and test splits, each for the respective learning scheme. Types of anomalous behavior are divided into group activities (e.g., Fighting, Theft, Slapping, Playing with a Racket, …) and individual activities (e.g., Throwing, Running, Littering, Sleeping, …) resulting in 22 classes of anomalous behavior. Anomalous behavior is frame-level annotated by hand, therefore the starting point and the endpoint

of the anomalous action is marked and frames in between are considered as anomalous. The dataset does not provide pixel-level anomalous annotations, which makes it difficult to localize the anomalous action. Activities that are not included are considered as normal (e.g., Walking, Waving, Talking, etc.). CHAD has 13 actors with diverse demographics (gender, age, ethnicity, etc.) which are appearing both in normal and anomalous clips.

CHAD is the first dataset to include annotations on individuals, which has two advantages:

(i) SBAD methods have access to processed data without the need for costly extraction of the data itself, making SBAD more accessible to other researchers.

(ii) It Provides a more unified approach and a comparable benchmark dataset that evaluates the skeleton-based approaches to anomaly detection without the pre-processing step of extracting human keypoints.

Therefore, the object detector YOLO (in particular YOLOv4) [Boc20] is utilized for extracting the human bounding boxes. These bounding boxes are then used for the person identifier labels, by feeding them into the DeepSort algorithm [Woj17], which requires a three frame warm-up (first two frames of each video). This results in unique person identifier labels for each person in a CHAD clip. The human poses are composed of human keypoints, thereby CHAD follows the skeleton model proposed by Lin et al. [Lin15] and introduced earlier in this thesis in Section 2.2. These keypoints are extracted using HRNet [Sun19] with a person confidence threshold of 0.5 (i.e., at least 9 keypoints of the person are below the threshold). Due to the removal of human poses, CHAD compensates this problem with the so-called *annotation smoothing*. Annotation smoothing uses high confidence annotations and linear interpolation to fill in the missing detections of either keypoints or bounding boxes [Dan23].

In summary, CHAD provides consistent guidelines for any skeleton-based approach to anomalous behavior detection due to its self-generated person annotations, high resolution video, and the fact that it is a more challenging

dataset than others [Paz22]. However, CHAD does not provide pixel-level annotations for anomalies, is rather utilized for unsupervised training and the annotations of human keypoints and bounding boxes may be outdated in the future.

5.3.3.1.4 Cannstatter Wasen

The CaWa18 (Cannstatter Wasen (CaWa) 2018) is an internal, hence not publicly available dataset of the Fraunhofer IOSB and part of the CrowDPB dataset. The subset was created in 2018 and consists of one recorded video from a static camera showing a crowded scene. This scene films the festival area of the annual "*Oktoberfest*" at the so-called *Cannstatter Volksfest* in the German city of Stuttgart, to be more precisely in Bad Cannstatt a borough of Stuttgart.

The video was recorded in 1080p at 25 frames per second with a total number of 306 frames which were filmed from an elevated position, to represent a surveillance video setting in a real-world crowded scenario.

The dataset does not contain any anomaly labels and is therefore only used for evaluation of the different pre-trained models on the various trainable datasets, to evaluate the performance of these on real-world scenarios. This dataset presents unique challenges due to the complexity of scenes, with each frame featuring more than 100 festival attendees. This results in numerous instances of occluded and self-occluded individuals, as well as various objects such as a balloon selling stand, which partially obstruct the view of different areas in the scene.

5.3.3.1.5 Hafengeburtstag Hamburg

Along with the CaWa18 dataset, HGH18 (Hafengeburtstag Hamburg (HGH) 2018) is another subset of CrowDPB and is not publicly available. Like the CaWa18 dataset presented in the preceding section, the dataset was recorded in 2018 and contains one sequence. This snapshot captures a scene within the Hafengeburtstag Hamburg festival in the German city of Hamburg. It encompasses a pedestrian crossing bridge from the middle right to the top left,

a street stretching from the middle left to the top right beneath the bridge, all situated to the right of the festival area.

The video shares the same specs as the CaWa dataset, i.e., has been recorded in 1080p at 25 frames per second. It consists of a total number of 800 frames, which were filmed from an very high position to overlook the pedestrian bridge and parts of the festival. Here is to be mentioned that the pedestrians that appear in the video are very small in scale as the camera position is way higher than in any other dataset. The camera is in such a high position to get a greater view on the pedestrian bridge, street and festival area. For future reference, we restrict the utilization of this dataset to frame indices up to 745, as subsequent frames are exclusively labeled every fifth frame with manually annotated bounding boxes. The dataset does not contain any anomaly labels and is therefore only used for evaluation of various pre-trained models on the different trainable datasets, to see the effectiveness of these on this real-world scenario.

5.3.3.1.6 Hessischer Rundfunk

The HeR19 dataset is another subset of CrowDPB that is not publicly available and includes two video clips provided by the Hessischer Rundfunk (HeR). Both sequences depict the same perspective with two main characters, a male and a female, encountering each other in a parking lot. Both videos depict the male approaching the woman, leading to a fight in the first video and an overdone hug in the second.

This subset focuses on highlighting the difficulties associated with analyzing behavior in natural environments.

The recordings were shot in 1080p resolution at a frame rate of 25 frames per second. In total, there are 1,277 frames captured from an elevated location commonly utilized for surveillance purposes. The dataset lacks anomaly labels and is solely utilized to assess the performance of several pre-trained models on different trainable datasets, in order to evaluate their effectiveness in a real-world scenario.

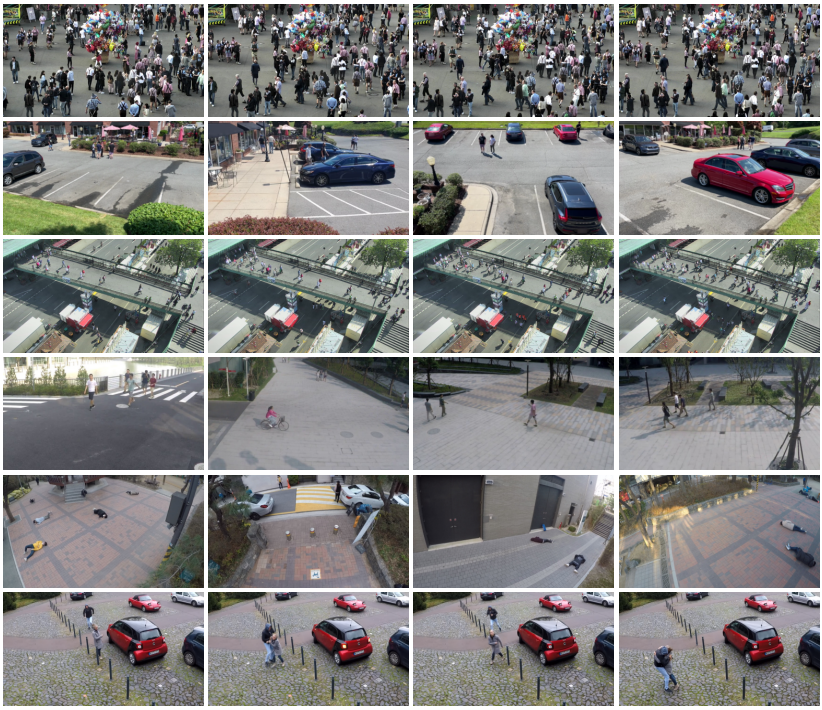


Figure 5.5: Exemplary frames from the different presented datasets. Each row contains four images per dataset, from top to bottom: CaWa18, CHAD, HGH18, SHTC-HR, VFP290K, and HeR19.

5.3.3.1.7 Summary of Datasets

In summary, all datasets were carefully selected to closely and accurately represent relevant surveillance scenarios. For this thesis, experiments were conducted to evaluate behavior analysis on six datasets. Three public datasets are used for quantitative comparison with existing research, while the other three are internal solely used for qualitative evaluation. The information is summarized in Table 5.4. Figure 5.5 shows exemplary frames from each of the introduced datasets.

Table 5.4: Summarized overview of the most relevant specs of the in Section 5.3.3.1 presented datasets. Note that VFP290K does not provide any distinct views since it was reecorded with a non-static camera.

Dataset	Frame Res.	fps	#Scenes	#Cameras	Type
SHTC-HR	480p	24	13	13	quan.
VFP290K	1080p	25	131	49	quan.
CHAD	1080p & 720p	30	13	4	quan.
CaWa18	1080p	25	1	1	qual.
HGH18	1080p	25	1	1	qual.
HeR19	1080p	25	2	1	qual.

5.3.3.2 Evaluation Pipeline

This subsection gives a brief overview of the pipeline proposed for the experiments performed in this work. It also provides implementation details for the object detection task performed with YOLOv8 [Joc22], the extraction of poses using HRNet [Sun19], as well as details for tracking individuals using the StrongSORT [Du23] tracker. The overall pipeline is visually presented in Figure 5.6.

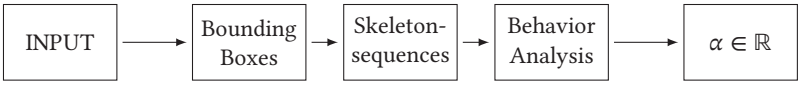


Figure 5.6: Illustration of a semi-automatic dataset preparation pipeline. Its aim is to provide a unified way to compare skeleton-based behavior analysis modules and generally rate their real-world capability.

Starting point is the raw data, i.e., the images in the datasets. These are utilized for person detection, as well as for the extraction of poses and the re-identification (ReID) of individuals. With respect to pedestrian annotations for specific frames, the datasets are divided into two categories: Datasets that already contain bounding boxes revisited by humans include the VFP290K, CaWa18 and HGH18 datasets. On the other hand, datasets that do not contain human-revisited bounding boxes for persons appearing in the images include the CHAD, ShanghaiTech Campus (SHTC), and HeR19 datasets. In order to

evaluate on the latter, the YOLOv8 [Joc22] persond and object detector, together with the Slicing Aided Hyper Inference (SAHI) [Aky22] approach for the HeR19 and CHAD datasets, is utilized to generate such. SAHI provides an inference tool developed mainly for the issue that recent object detectors excel in detecting large objects in low-resolution images from common datasets like ImageNet [Den09] and COCO [Lin15]. However, Akyon et al. [Aky22] state that they struggle with accuracy when tasked with detecting small objects in high-resolution images produced by advanced drones and surveillance cameras. The slicing method involves dividing the original image into overlapping patches for independent object detection. The results are merged using Non-maximum Suppression (NMS) to the original size, wherein boxes with IoU ratios above a pre-defined threshold are adopted, and detections with probabilities below another threshold are removed. In this work, SAHI is configured to have an overlap ratio of patches by 30% in terms of height and width, which are similar to the values used in the originating approach. The size of the patches is defined as 640×640 for both the CHAD and the HeR19 dataset. These sizes were chosen based on the largest size of appearing humans within the videos of the dataset, with respect to the time complexity of the multiple performed object detection tasks due to this approach. Subsequently, the large version of the YOLOv8 model, pre-trained on the COCO dataset with an input resolution of 640×640, is utilized for the datasets CHAD, SHTC and HeR19. Note that only the person class is included in this work, while all other detections are discarded. The goal is to achieve high-quality person detections for the object detection task, resulting in human bounding boxes.

Subsequently, both aforementioned types of human bounding boxes serve as input for the human pose estimator HRNet and the tracker StrongSORT. In the case of predicting the joints of individuals within these human bounding boxes, this process selects the larger version of HRNet, denoted as HRNetW48, with input bounding box sizes of 384×288 for all datasets. The HRNet, pre-trained on the training set of the COCO dataset, performs MPPE, resulting in semantic pose graphs with 17 keypoints (cf. Figure 2.5). The multi-object tracker StrongSORT provides identity labels, which utilizes the bounding boxes and the corresponding parts of the image to label the corresponding individuals with such identity labels over time. For the person

ReID process, the lightweight deep ReID model Omni-Scale Network (OSNet) [Zho19] is applied. All other hyperparameters are chosen as introduced in [Du23]. Note that only individuals with object detection confidences above 50% are tracked by OSNet, which is a limiting factor in respect to the construction of human pose trajectories. For bounding boxes annotated by humans confidences are set to 100% to avoid them being filtered out.

Due to the usage of recently published approaches such as the YOLOv8 object detector, HRNet human pose estimator and the StrongSORT tracker, this work achieves extensive and qualitative input samples for the SBAD models in the subsequent analysis.

5.4 Experimental Results

This section presents the results of various experiments conducted to investigate the performance of the developed approaches. Firstly, following the order as the previous sections, Section 5.4.1 presents the results of the experiments for the task of DA as introduced in Section 5.3.2. Secondly, Section 5.4.2 does so for the task of behavior analysis. The focus lays here on the microscopic SBAD approach, where the macroscopic one is taken just as a baseline Section 5.3.3. Part of both sections are quantitative as well as qualitative evaluations.

5.4.1 Domain Adaptation

5.4.1.1 Qualitative Evaluation of Domain Adaptation

This section takes a look on the qualitative results of the domain adaptation process with respect to the two datasets that were chosen as target domain, namely WorldExpo '10 in the first part of the section, followed by HGH18-DA. Both datasets were selected to fit the source domain as close as possible.

Figure 5.7: Exemplary frames from WorldExpo '10 dataset. Apparently the overall image quality is poor and various artifacts, like the black uneven border around every frame, selective blur due to smudgy lenses or the blurred section in the lower right corner, which might have been originally a certain kind of text overlay or timestamp.

5.4.1.1.1 WorldExpo '10

The first target domain that is considered for adapting SyMPose into is defined by the World Expo (WE) dataset. The images contained within the dataset are of rather low quality, which comes from various blurred regions and image artifacts. However, as stated in [Gol20] its setting is close to the source domain many people depicted in the images, which is why it is of interest to evaluate the DA on this dataset. At first glance, the U-Cycle-GAN approach produces convincing results for the dataset as target domain, some of which can be observed in Figure 5.8.

Figure 5.8: Exemplary results of the DA process from SyMPose to WorldExpo '10. Obviously the general appearance of the visualized frames unveils a certain similarity to the original images shown in Figure 5.7.

The images are sharp and contain nearly all of the content details of the synthetic images from the source domain. Despite the aforementioned artifacts in the target domain, the adapted images show a comparable quality to those from the source images which seems to be desirable, from a visual point of view. However, this preservation of image quality could be an issue when utilizing these images as training data as a detection model that was trained on such adapted data would probably face challenges when being applied to the target domain data. This should be kept in mind when looking at these images. Some artifacts that occur regularly in the domain adapted data samples are shown in Figure 5.9.

Figure 5.9: Exemplary artifacts from WE-adapted dataset. Despite the overall promising appearance, there can be found several reoccurring kinds of artifacts. However, each of these are typically only affecting small regions and as long they do not affect pedestrians, they will not play an important or even any role for the HPE task.

The most often occurring artifacts are various "color burns" that happen in arbitrary places. As most of them do not occur on the shown pedestrians they can be seen as potentially harmless to the following training process, since the HPE approach that is trained on this data later on will not encounter most of these artifacts.

5.4.1.1.2 Hafengeburtstag Hamburg 2018

The second target domain that is taken into respect is the HGH18-DA which has been presented in Section 5.3.2.1. Compared to the WE dataset, the overall quality of the image material provided by the HGH18-DA is good. One

Figure 5.10: Exemplary results of the DA process from SyMPose to HGH18. Different to the results obtained for WE the frames show an overall dark appearance, stronger contrast and bluish tint. Furthermore, various artifacts in the domain adapted images are present as well, although very similar to those in the WE-adapted frames.

key difference is the overall resolution of the resulting frames, which were recorded with a state-of-the-art IP camera. As already mentioned for the WE dataset the U-Cycle-GAN produces overall convincing results for the DA task, as illustrated in Figure 5.10. In general the same observations that were made for the first dataset can be made for the second one as well: the details of the source domain data is preserved in most cases. Apparently, the network has learned to give its input images a cyan tint based on the target domain data. This might come from an significant amount of samples showing the copper roof at the St. Pauli Piers in Hamburg. Some exemplary generated patches are illustrated in Figure A.10. Something that is conspicuous, are the yellow areas on the street that happen in particular on crosswalks. This looks, especially in the case of wet roads to the appearance of light reflections from low sun, however this is something that is not represented in the target domain data. This leads to the assumption, that this particular pattern in the source domain cannot be handled appropriately by the model. However, there are even more artifacts as is illustrated in Figure 5.11. All in all the adapted image material shows a darker appearance, with strong contrast and the bluish tint, all being properties that distinguish the data from the WE-adapted ones. As

reported in [Gol20] this does not have to be necessarily a bad sign, it might be even beneficial to the training process of an HPE approach. The artifacts as observed in the qualitative analysis may play an important role as they can be seen as artificial occlusions that can have a regularizing character during training since the pose annotation is still available.

Figure 5.11: Exemplary artifacts from HGH18-adapted dataset. As mentioned in Figure 5.10 the samples generated by adapting SyMPose to HGH18 appear to be worse than those adapted to WE. The most prominent types are the color of the cyan tint of the sky and cert "burn outs" as shown on the right.

To summarize, both DA processes can be seen as successful, at least from a visual perspective. The next section dives into the quantitative experiments and investigates the effects on the training of a HPE approach.

5.4.1.2 Quantitative Evaluation of Domain Adaptation

In order to assess the performance of HPE approaches trained on synthetic and domain adapted data, different models were trained on the generated data to investigate the impact on the training. Although, Fréchet inception distance (FID) is a widely used metric to assess the quality of image generation algorithms, according to Chong et al. [Cho20] it is not suited to evaluate domain adaptation tasks. Therefore the comparison of the different datasets and models is based on the evaluation procedure presented in Section 5.3.2.3.

5.4.1.2.1 Comparison of Different Human Pose Estimation Models

First and foremost various state-of-the-art HPE models are compared in order to find an appropriate choice for the following experiments. The set of

Table 5.5: Model comparison on CrowdPE for four different models: SimpleBaselines [Xia18], HRNetW32 [Sun19], HRNetW48 [Sun19], and ViTPose-B [Xu22b]. SimpleBaselines has been used in [Gol20], while the others represent state-of-the-art methods. The models were trained on CrowdPose [Li19] (upper part) and SyMPose (lower part).

Model	mAP	mAR	F1
SimpleBaselines	0.232 ± 0.004	0.262 ± 0.003	0.247 ± 0.004
HRNetW32	0.171 ± 0.059	0.227 ± 0.038	0.194 ± 0.054
HRNetW48	0.227 ± 0.003	0.261 ± 0.003	0.243 ± 0.003
ViTPose-B	0.187 ± 0.003	0.224 ± 0.003	0.203 ± 0.003
SimpleBaselines	0.164 ± 0.002	0.193 ± 0.001	0.177 ± 0.002
HRNetW32	0.179 ± 0.008	0.212 ± 0.007	0.194 ± 0.008
HRNetW48	0.188 ± 0.005	0.221 ± 0.004	0.203 ± 0.005
ViTPose-B	0.117 ± 0.009	0.148 ± 0.007	0.131 ± 0.008

models consists of the SimpleBaselines [Xia18] model that was initially used in [Gol20], as well as the HRNetW32 [Sun19], HRNetW48 [Sun19], and ViTPose-B [Xu22b].

To find an appropriate choice, the models were trained on real-world as well as synthetic data. For the above mentioned models real-world trainings were conducted on CrowdPose [Li19], whereas the trainings on synthetic data were performed on the introduced SyMPose dataset. To compare these trained models, they are evaluated on CrowdPE as SimpleBaselines, which is the only model that was not trained along with the other models, only reports results on the mentioned dataset. As explained in Section 5.3.1, each image of the evaluation dataset is characterized with respect to its crowdedness using either the CI or sGCI. This characterization allows a partitioning of the dataset into easy, medium and hard samples, following the procedure initially proposed by Li et al. [Li19] and adapted in [Gol20]. The consequent results are reported in Table 5.5 for both, the real-world and the synthetic trainings, in the upper and lower part of the table, respectively. Apparently, the models trained on real-world data perform better than those trained on the synthetic samples, which can be observed over all provided metrics. Despite being a result to be expected, the gap between both domains diminishes going from easy to medium. As the Tables 5.5 and 5.6 also indicate, there is a significant difference between the different models. While SimpleBaselines achieves the

Table 5.6: Analogous to Table 5.5 this table provides results for different models trained on CrowdPose [Li19] (upper part) and SyMPose (lower part). The table gives an insight into how overall performance distributes over the different categories for both, the CI- and sGCI-based splits.

	CI		sGCI		
Model	**mAP [e]**	**mAP [m]**	**mAP [e]**	**mAP [m]**	**mAP [h]**
SimpleBaselines	0.773 ± 0.016	0.224 ± 0.003	—	—	—
HRNetW32	0.679 ± 0.022	0.166 ± 0.058	0.443 ± 0.122	0.211 ± 0.057	0.072 ± 0.054
HRNetW48	0.696 ± 0.023	0.221 ± 0.003	0.529 ± 0.017	0.258 ± 0.003	0.127 ± 0.004
ViTPose-B	0.547 ± 0.020	0.183 ± 0.003	0.400 ± 0.013	0.224 ± 0.004	0.089 ± 0.003
SimpleBaselines	0.451 ± 0.029	0.161 ± 0.003	—	—	—
HRNetW32	0.456 ± 0.124	0.176 ± 0.007	0.338 ± 0.042	0.215 ± 0.007	0.089 ± 0.007
HRNetW48	0.558 ± 0.035	0.184 ± 0.005	0.377 ± 0.022	0.225 ± 0.006	0.091 ± 0.004
ViTPose-B	0.109 ± 0.056	0.117 ± 0.009	0.117 ± 0.030	0.153 ± 0.014	0.050 ± 0.005

best results of all models when trained on real-world data, its performance decreases significantly when training on the synthetic SyMPose dataset. The other models achieve a more consistent performance over both scenarios, however HRNetW32 occasionally fails to converge during training on both domains and ViTPose-B performs exceptionally poor as it seems to struggle with the overall synthetic domain.

To summarize, all of the shown models achieve mediocre performance on the surveillance scenario that is represented by the evaluation dataset CrowdPE. This shows how challenging the given scenario is, especially keeping in mind that the HPE task is relevant to behavior analysis.

For the following experiments the model of choice is HRNetW48 for two reasons: First, as the results indicate HRNetW48 shows comparable performance to the other models when trained on real-world data and outperforms most of the other models when trained directly using synthetic data. Second, the choice keeps this work consistent with the respect to the second part of the thesis, where the HRNetW48 is also used. All in all, HRNetW48 appears to be the best choice and an appropriate trade-off between performance on synthetic and real-world data.

Note that the results for SimpleBaselines as given in the Tables 5.5 and 5.6 are those reported in [Gol20], which is also the reason that only results for

Table 5.7: Results of experiments using HRNetW48 trained on various datasets. All values are reported on CrowdPE. The first two rows show trainings on non-adapted data, namely the real-world dataset CrowdPose and the synthetic dataset SyMPose. The remaining rows show the results obtained on adapted data from SyMPose.

Dataset	Comb.	mAP	mAP [e]	mAP [m]	mAP [h]
CrowdPose	—	0.227 ± 0.003	0.529 ± 0.017	0.258 ± 0.003	0.127 ± 0.004
SyMPose	—	0.188 ± 0.005	0.377 ± 0.022	0.225 ± 0.006	0.091 ± 0.004
WE	—	0.152 ± 0.006	0.221 ± 0.023	0.188 ± 0.010	0.076 ± 0.006
HGH18	—	0.168 ± 0.006	0.311 ± 0.017	0.205 ± 0.006	0.081 ± 0.005
WE + HGH18	*merge*	0.180 ± 0.008	0.334 ± 0.033	0.219 ± 0.010	0.084 ± 0.005
WE + HGH18	*full*	0.175 ± 0.002	0.324 ± 0.020	0.213 ± 0.006	0.083 ± 0.003

the CI-based dataset splits are given and none for the sGCI-based partitions. Furthermore, in general any kind of HPE network can be used for the evaluation procedure. This specific selection is based on personal experience with certain of these models, the overall performance on public benchmarks, and the availability of open source high-quality source code.

5.4.1.2.2 Impact of Domain Adapted Data on the Training

From this point on the reported results are exclusively obtained using the HRNetW48 model. Subject to this section are experiments using domain adapted data that are obtained using the U-Cycle-GAN architecture presented in Section 4.3.3 and following the evaluation procedure described previously. For convenience, each table provides the results obtained when training the model on *pure* SyMPose and CrowdPose as reference.

Table 5.7 displays the results obtained from training the HRNetW48 on data that was adapted to the WorldExpo '10 and the HGH18 dataset, including those trained on the bare synthetic and real-world data. The first two experiments shown in Table 5.7 are each trained solely on one of the two introduced domains. First of all, it can be stated that the performance of the HRNetW48 deteriorates slightly due to the training on domain-adapted data. This can be observed over all three categories and is hence reflected by the overall performance as well. Especially the model that was trained on the WorldExpo

Table 5.8: Additional mAR and F1 scores extending Table 5.7.

		CrowdPE	
Dataset	**Comb.**	**mAR**	**F1**
CrowdPose	—	0.261 ± 0.003	0.243 ± 0.003
SyMPose	—	0.221 ± 0.004	0.203 ± 0.005
WE	—	0.184 ± 0.005	0.166 ± 0.006
HGH18	—	0.202 ± 0.005	0.183 ± 0.005
WE + HGH18	*merge*	0.212 ± 0.007	0.194 ± 0.007
WE + HGH18	*full*	0.208 ± 0.002	0.190 ± 0.002

'10 data drops in recognition performance and shows inferior performance not only compared to the reference models but also with respect to the model trained on HGH18-adapted data. One particular reason for the bad performance of the model that was trained on WorldExpo '10 data is, that HGH18 is visually closer to the evaluation dataset CrowdPE, since the dataset contains examples recorded at the same place in Hamburg. Yet these were collected some years earlier with a slightly different hardware setup. However, the properties of the HGH18-adapted data seem to be a better fit than those from WorldExpo '10.

Combining Datasets. As stated before, the two target domains for the DA differ from the evaluation dataset, resulting in a dataset dependent domain shift between adapted and evaluation data. Consequently, none of the two adapted datasets would be optimally matching the whole evaluation dataset, even if the adapted images were exactly alike the target domain. One way to address this issue is to join the datasets that were used so far in order to increase the generalization abilities of the trained model as a better generalizing model should also perform better on the evaluation dataset.

The adapted datasets are therefore joined in two ways with each other as well as with the synthetic data,

(i) such that the resulting joined set contains every image of the original source domain once,

(ii) such that for each sample of the original source domain all the corresponding target samples are included.

The first case (i) can be formulated more formally as follows: Given the source dataset X (i.e., SyMPose), with $|X| = N$ a split of X into k distinct subsets $X_i \neq \emptyset \forall i \in \{0, \dots, k-1\}$ with

$$X = \bigcup_{i=0}^{k-1} X_i \tag{5.14}$$

and $\forall i, j \in \{0, \dots, k-1\}$ with $i \neq j$

$$X_i \cap X_j = \emptyset. \tag{5.15}$$

So far this defines a partition of X. For the combination of datasets, this is extended by another property, which ensures that all subsets are of equal size:

$$|X_0| = |X_1| = \cdots = |X_{k-1}| \tag{5.16}$$

The joined dataset is then defined as $\sum_{i=0}^{k-1} \tilde{X}_i = \bigcup_{i=0}^{k-1} \tilde{X}_i$, with transformation functions $f_i \forall i \in \{0, \dots, k-1\}$ and $\tilde{X}_i := f_i(X_i)$ that map every element from X_i to the corresponding sample of the target set. Note that obtaining the same cardinality for all X_i cannot always be ensured since the number of elements contained in X does not have to be divisible by k. In such cases, the cardinality of the subsets differs slightly, which however is negligible and will not have an effect on the overall training performance.

Creating the training sets for the second case (ii) is much easier. Here, the datasets are just put together completely, i.e., every sample from the corresponding dataset is included resulting in a training set with $k \cdot N$ samples, whereas the first case will by definition always result in datasets with N samples. In the experiment tables, these two scenarios are referred to as *merge* and *full* for (i) and (ii), respectively.

Table 5.7 furthermore reports the detection results of those models that were trained on these joined datasets. As expected, the resulting outcomes are better than those presented before underlining the increase of generalization that can be achieved using such combined data samples. The performance increases over all categories, which implies that the trained model generalizes

Table 5.9: Results of experiments using HRNetW48 trained on various datasets including samples from the synthetic source dataset SyMPose. All values are reported on CrowdPE. The first two rows show trainings on non-adapted data, namely the real-world dataset CrowdPose and the synthetic dataset SyMPose. The remaining rows show the results obtained on different combinations of adapted and synthetic data, either merged or fully combined.

Dataset	Comb.	mAP	mAP [e]	mAP [m]	mAP [h]
CrowdPose	—	0.227 ± 0.003	0.529 ± 0.017	0.258 ± 0.003	0.127 ± 0.004
SyMPose	—	0.188 ± 0.005	0.377 ± 0.022	0.225 ± 0.006	0.091 ± 0.004
syn. + WE	*merge*	0.190 ± 0.006	0.374 ± 0.017	0.228 ± 0.007	0.093 ± 0.006
syn. + HGH18	*merge*	0.192 ± 0.003	0.339 ± 0.017	0.233 ± 0.006	0.095 ± 0.002
syn. + WE + HGH18	*merge*	0.192 ± 0.004	0.370 ± 0.028	0.233 ± 0.007	0.094 ± 0.006
syn. + WE	*full*	0.188 ± 0.003	0.366 ± 0.018	0.226 ± 0.005	0.093 ± 0.004
syn. + HGH18	*full*	0.190 ± 0.002	0.368 ± 0.029	0.230 ± 0.002	0.089 ± 0.005
syn. + WE + HGH18	*full*	0.191 ± 0.004	0.383 ± 0.013	0.229 ± 0.006	0.092 ± 0.005

better. With respect to the combination type there seems be no clear difference with respect to detection performance. The *merge* scenario appears to perform better on average compared to the *full* scenario, yet at the same time it comes with an increased variance in the obtained results.

The combination of datasets does not have to be limited to the domain adapted sets. Since the synthetic SyMPose data, i.e., the non-adapted data, is also available, it can be used as well to enrich the training sets even more. This is subject to the results shown in Table 5.9, where analogous to the experiments in Table 5.7 each training set consists of multiple parts. Evidently, the additional usage of the pure unadapted data diminishes the variance of the results of the combined datasets and increases overall the performance. However, different than assumed, combining the pure synthetic data with the domain adapted ones does not seem to increase the generalization performance of the HRNetW48. Similar to the observations made for the merges between WE- and HGH18-adapted, the *full* scenario reduce the variance in most cases. The reason for this can be found in the multiple visual variants of every pedestrian that are included in the final training datasets. Since the original pose annotation are not altered by the domain adaption process, for each sample there are up to three different visual representation. Being in some cases sharper

Table 5.10: Additional mAR and F1 scores extending Table 5.9.

		CrowdPE	
Dataset	**Comb.**	**mAR**	**F1**
CrowdPose	—	0.261 ± 0.003	0.243 ± 0.003
SyMPose	—	0.221 ± 0.004	0.203 ± 0.005
syn. + WE	*merge*	0.221 ± 0.005	0.205 ± 0.005
syn. + HGH18	*merge*	0.224 ± 0.002	0.207 ± 0.002
syn. + WE + HGH18	*merge*	0.224 ± 0.002	0.207 ± 0.003
syn. + WE	*full*	0.220 ± 0.003	0.203 ± 0.003
syn. + HGH18	*full*	0.222 ± 0.004	0.204 ± 0.002
syn. + WE + HGH18	*full*	0.223 ± 0.004	0.206 ± 0.004

and in other more blurry as well as different color appearances may have an regularizing effect on the overall training.

Comparison on HGH18. To address the issue of dataset dependent domain shift that was mentioned above, additional experiments are carried out solely using the HGH18-HPE data introduced in Section 5.3.1. The results are depicted in Tables 5.11 and 5.12. As is immediately apparent, the resulting scores are significantly lower than those obtained on CrowdPE. Even the model trained on the real-world dataset CrowdPose performs poorly, which underlines the challenging character of the scenario defined by the HGH18 dataset. The main reasons for this observation is that the HGH18 dataset consists of a large number of people with a particularly low resolution, as is exemplarily illustrated in Figure 5.12. Different than in CrowdPE ($\bar{\mu} = 120.0$, $\bar{\sigma} = 68.7$) these pedestrians have approximately the same height ($\bar{\mu} = 60.9$, $\bar{\sigma} = 15.1$) since all the samples of the dataset come from one scene, whereas CrowdPE was collected in various scenarios using different camera setups. As the example images of pedestrians already show, most of the difficulty originates from the aforementioned aspect, since mutual occlusions between pedestrians occur less often than in CaWa18. Since the evaluation dataset HGH18-HPE and the DA dataset HGH18-DA share the same origin, the scenario describes an ideal case as both domains should be almost identical or at least barely distinguishable. Analogous to the previous tables, Tables 5.11 and 5.12 provide two additional rows that show the performance of the HRNetW48 trained on real-world and synthetic data as reference. The remaining

Figure 5.12: Pedestrian examples from HGH18-HPE. The examples show the area around random pedestrians with the target person being right in the center of the area. All people within the dataset have in common, that they are of low resolution, which makes it especially challenging to estimate their poses. For this particular dataset this is the most challenging property, as mutual occlusions with other pedestrians occur less often compared to CaWa18.

table is divided into three parts, the trainings on the two bare adapted datasets, as well as the two types of dataset combinations. First of all, the models that were trained exclusively on either WE or HGH18 show a larger variance with respect to the mAP compared to the two reference cases. From the outcomes of these two experiments it seems that the training on WE-adapted data leads to decrease in the mAP, while increasing the mAR at the same time. HGH18-adapted data however increases on average both, the mAP and mAR. This is the anticipated result, and indicates that the adaptation of the synthetically generated data was successful. However, the variance of the obtained outcomes is still higher than those of the reference cases, which might come from the artifacts that were addressed in the qualitative evaluation presented in Section 5.4.1.1. This changes when combining different datasets, which reduce the aforementioned variance again, and especially with the HGH18-combined datasets even an increase of the performance can be achieved.

The results show that given a well-defined scenario it is possible to improve the performance of a HPE model by using synthetically generated data that was adapted to the targeted domain. In the particular case defined by video surveillance in urban setups, the experiments revealed that estimating poses of pedestrians is still a very challenging task that requires even further work and offers space for improvement. With regard to the performance on the HGH18 dataset, two assumptions arise: Firstly, the quality of the data representing the target domain that has a direct influence on the quality of the

Table 5.11: Comparison of all experiments conducted on HGH18 and CaWa18. Since both evaluation datasets fall mostly into the categories *medium* and *hard*, the comparisons solely refers to the overall mAP values on both datasets.

Dataset	Combination	mAP @HGH18	mAP @CaWa18
CrowdPose	—	0.055 ± 0.003	0.103 ± 0.001
SyMPose	—	0.050 ± 0.004	0.087 ± 0.001
WE	—	0.045 ± 0.013	0.085 ± 0.003
HGH18	—	0.058 ± 0.011	0.085 ± 0.002
WE + HGH18	*merge*	0.057 ± 0.008	0.086 ± 0.004
syn. + WE	*merge*	0.051 ± 0.006	0.087 ± 0.001
syn. + HGH18	*merge*	0.066 ± 0.005	0.087 ± 0.002
syn. + WE + HGH18	*merge*	0.065 ± 0.007	0.087 ± 0.001
WE + HGH18	*full*	0.051 ± 0.008	0.089 ± 0.001
syn. + WE	*full*	0.059 ± 0.006	0.087 ± 0.001
syn. + HGH18	*full*	0.065 ± 0.006	0.086 ± 0.003
syn. + WE + HGH18	*full*	0.065 ± 0.005	0.086 ± 0.002

Table 5.12: Additional mAR and F1 scores extending the mAP results from Table 5.11 on HGH18 and CaWa18.

		HGH18		CaWa18	
Dataset	**Comb.**	**mAR**	**F1**	**mAR**	**F1**
CrowdPose	—	0.089 ± 0.002	0.068 ± 0.003	0.105 ± 0.001	0.104 ± 0.001
SyMPose	—	0.081 ± 0.002	0.062 ± 0.004	0.090 ± 0.001	0.088 ± 0.001
WE	—	0.078 ± 0.010	0.057 ± 0.013	0.089 ± 0.002	0.087 ± 0.003
HGH18	—	0.088 ± 0.009	0.070 ± 0.011	0.090 ± 0.001	0.087 ± 0.002
WE + HGH18	*merge*	0.089 ± 0.008	0.070 ± 0.008	0.091 ± 0.003	0.088 ± 0.003
syn. + WE	*merge*	0.088 ± 0.004	0.066 ± 0.005	0.090 ± 0.001	0.089 ± 0.001
syn. + HGH18	*merge*	0.096 ± 0.004	0.078 ± 0.004	0.091 ± 0.002	0.089 ± 0.002
syn. + WE + HGH18	*merge*	0.097 ± 0.004	0.077 ± 0.006	0.091 ± 0.001	0.089 ± 0.001
WE + HGH18	*full*	0.086 ± 0.007	0.064 ± 0.008	0.092 ± 0.001	0.090 ± 0.001
syn. + WE	*full*	0.092 ± 0.005	0.072 ± 0.006	0.091 ± 0.001	0.089 ± 0.001
syn. + HGH18	*full*	0.094 ± 0.007	0.077 ± 0.007	0.091 ± 0.001	0.088 ± 0.002
syn. + WE + HGH18	*full*	0.097 ± 0.005	0.078 ± 0.005	0.091 ± 0.002	0.089 ± 0.002

domain-adapted data. This process may need a even more careful collection of data. Secondly, the eligibility of the source data, i.e., SyMPose, itself may be an issue as well. For example, although SyMPose was designed for the exact scenario, it may still have too much variance in its visual appearance, but at

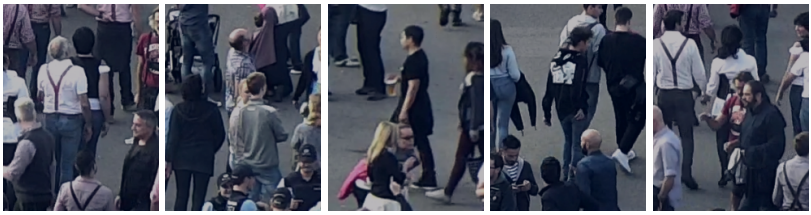


Figure 5.13: Pedestrian examples from CaWa18. Analogous to Figure 5.12, the examples show the area around random pedestrians with the target person being in the center of the area. Different to the HGH18-DA dataset, the people are larger, but at the same time there are much more mutual occlusions as the examples indicate.

the same time too little variance in its provided human poses. Even the very limited number of pedestrian models and their *virtual* ethnicity may have an negative effect on the performance. As the examples given in Figure A.4 indicate, there is a relatively high number of dark-skinned people, which is most probably based on the US-American society. This may be a limiting factor for the final performance of the trained models with respect to the used datasets, which in return are closer to the appearance of the Central European.

Comparison on CaWa18. As mentioned earlier, the CaWa18 is quite different to the HGH18 dataset, since the overall pedestrian sizes are significantly larger ($\bar{\mu} = 188.3$, $\bar{\sigma} = 59.1$) as the examples illustrated in Figure 5.13 signify. At the same time the scene is much more crowded resulting in more frequently occurring mutual occlusions which increase the difficulty of extracting correct poses. The results that were obtained on CaWa18 are contained in Tables 5.11 and 5.12 as well. As it was to be expected due to the performance on CrowdPE, the models achieve higher scores than on the HGH18 dataset which is possibly related to the crisper and higher-quality images of the pedestrians. Since none of the two chosen target domains WorldExpo '10 and HGH18 is a real close fit, the CaWa18 is adduced for comparing the different trained models with regard to their generalization ability. The results indicate that a certain resolution is necessary to estimate robust human poses, however in the particular scenario the observations that were made for the HGH18 do not hold for the CaWa18 dataset. While combining the three available domains was beneficial for the performance on HGH18, it appears that the models do not

avail from data being combined as they seem to be too far from the targeted domain. It seems like there is an improvement compared to just training on the single domain-adapted datasets, but the overall increase in performance is marginal. Here the issue of having an appropriate source domain might be even more striking, since none of the datasets helped to close or reduce the gap between the two reference models. In this particular case, the scenes from SyMPose-Urban that were recorded with a short focal length, differ strongly from the CaWa18 sequence, which was evidently recorded with a far longer focal length. The results brings up the already stated assumption that the source domain data, i.e., SyMPose, may not be appropriately designed to use on scenarios like such represented by the CaWa18 dataset.

5.4.2 Behavior Analysis

5.4.2.1 Holistic Analysis: MurzGAN

As mentioned in the beginning of Section 5.4 the behavioral part focuses on the SBAD method, wheres the MurzGAN is mainly taken for comparison. However, the first experiments take a look at alternative backbones for the semantic feature extractor to investigate the choice of backbone and compare various state-of-the-art CNN- and Transformer-based models.

5.4.2.1.1 Performance Comparison of Semantic Backbones

To choose an appropriate feature extractor various pre-trained state-of-the-art models of the Swin Transformer (Swin) [Liu21c], ViT [Dos20], and ConvNeXt [Liu22b] along with the original VGG16 [Sim14] are compared. All mentioned architecture consist out of various layers or blocks, each of which generates a corresponding feature map. These different outputs are used to generate semantic feature vectors for the ground truth and generated frame of the GAN setup. Based on these semantic vectors, differences are computed, which are then used to assess the frame level performance of the different models. The summarized results are shown in Table 5.13 where for each model the results of the best performing layer or block output is displayed. Appar-

Table 5.13: Comparison of Semantic Backbones for MurzGAN. All reported scores are AUC-ROC on frame-level for the SHTC dataset. For the complete results of all model blocks see Tables A.1 and A.2.

Model	AUC-ROC
Swin-T	0.6049 ± 0.0877
Swin-B	0.6038 ± 0.0825
ViT-B16	0.6530 ± 0.1021
ViT-B32	0.6297 ± 0.0755
ConvNeXt-t	0.6030 ± 0.0959
ConvNeXt-s	0.5795 ± 0.0909
VGG16	0.6897 ± 0.0890

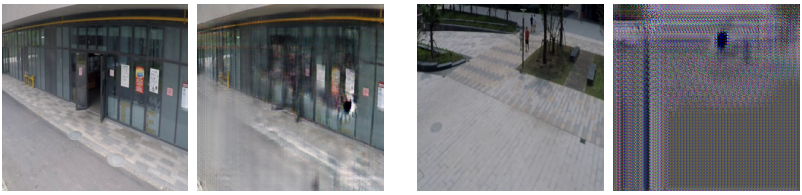


Figure 5.14: Two examples of the frame generation from the observed real optical flow. Each pair of images is a corresponding (real,fake) pair for a particular view of SHTC. The left setting is located closer to the camera, while the right shows a wider overview.

ently all model architectures that were examined achieve comparable results for the semantic feature extraction without achieving outstanding overall outcomes. It is particularly striking that the results obtained from the different models show a strong variance, which comes from the overall difficult GAN training. Depending on the scene, the MurzGAN struggles to converge and learn the scene as it is shown exemplary in Figure 5.14, which will naturally result in large semantic differences as the feature extractor used for the semantic features is not able to extract useful information. This is a behavior that can be observed in many cases and is especially crucial since a correct or at least very close generation of samples is indispensable for the recognition of anomalous behavior patterns. However, it is somehow expectable that the GAN would face certain difficulties, as the SHTC provides color information in contrast to the UCSD dataset, which is completely in grayscale. The colored images increase the difficulty for the GAN, as there is additional visual information that has to be generated beside the structures. In the main, this makes

Table 5.14: Inference speed performance evaluation per frame for different models on an Nvidia L40 based 8,192 runs per model. All reported times are given in milliseconds and correspond to the full model.

Model	Batch Size				
	1	**8**	**16**	**32**	**64**
Swin-T	6.65 ± 0.89	0.85 ± 0.13	0.83 ± 0.02	0.87 ± 0.02	0.92 ± 0.02
Swin-B	13.46 ± 1.81	2.19 ± 0.07	2.38 ± 0.04	2.30 ± 0.04	2.39 ± 0.03
ViT-B16	2.82 ± 0.37	2.27 ± 0.05	2.13 ± 0.04	2.11 ± 0.04	2.00 ± 0.04
ViT-B32	3.28 ± 0.60	0.54 ± 0.02	0.50 ± 0.03	0.54 ± 0.02	0.51 ± 0.01
ConvNeXt-t	2.91 ± 0.52	0.63 ± 0.02	0.66 ± 0.01	0.67 ± 0.02	0.75 ± 0.70
ConvNeXt-s	5.77 ± 6.23	1.12 ± 0.02	1.14 ± 0.01	1.21 ± 0.03	1.27 ± 0.30
VGG16	1.87 ± 0.07	0.85 ± 0.05	0.86 ± 0.07	0.90 ± 0.10	0.91 ± 0.13

it increasingly difficult to generate the correct corresponding frame, which will also result in larger semantic differences at the same time. Another difference between the UCSD dataset and the SHTC is the overall pedestrian size. The people in the scenes are much smaller for the UCSD and show less variance with respect to the aforementioned size. Although shown to work for easier and academic settings [Gol19c], the MurzGAN suffers from many aspects that make it inadvisable to use for real-world scenarios.

In addition to the bare recognition performance the models were furthermore examined with respect to their inference speed. Table 5.14 shows the inference speed per frame for the different semantic feature extractors with respect to different batch sizes. All architectures regarded here are capable of computing semantic features in real-time, since all values indicate frame rates between roughly 60 to 1,250 frames per second. The results underline in particular that more modern architectures do not necessarily offer faster inference, especially since the limiting factor for the MurzGAN is the GAN network. No further optimizations were made with regard to speed performance. The purpose of the study was primarily to support the decision-making process for selecting a suitable model. Finally, both experiments lead to the decision of keeping the initial VGG16 as the semantic backend since none of the examined models lead to a significant improvement for the MurzGAN. The next section takes a look on the overall performance of MurzGAN.

Table 5.15: Frame-level performance of MurzGAN on SHTC (top) and CHAD (bottom). Apparently MurzGAN achieves to distinguish anomalous from normal behavior on the SHTC dataset, while it fails to do so for CHAD as well.

Variant	AUC-ROC	AUC-PR	EER
Frame only	0.6552 ± 0.1404	0.6200 ± 0.1172	0.3883 ± 0.1058
Flow only	0.7070 ± 0.0920	0.7050 ± 0.2075	0.3151 ± 0.0750
Both	0.6762 ± 0.1225	0.6793 ± 0.1795	0.3632 ± 0.0926
Frame only	0.5109 ± 0.0230	0.4826 ± 0.0891	0.4981 ± 0.0406
Flow only	0.5051 ± 0.0157	0.4995 ± 0.0795	0.4561 ± 0.0156
Both	0.5194 ± 0.0181	0.4821 ± 0.0948	0.4850 ± 0.0354

5.4.2.1.2 Evaluation on SHTC and CHAD

As already indicated in the preceding section, MurzGAN struggles with real-world datasets like the SHTC. While the experiments on UCSD Pedestrian [Li13] dataset that were reported in [Gol19c] showed promising performance, the results on the two more challenging datasets SHTC and CHAD reveal that MurzGAN is not able to capture the relevant information to characterize a given situation sufficiently well. To be more concise, Table 5.15 shows results on both, SHTC and CHAD. For each of both datasets the performance using motion and appearance information as well as the combined information is displayed. First and foremost, the results in the lower part of the table show, that the approach is not able to achieve any positive results for the CHAD dataset as it performs almost like a random guessing. The model has neither the ability to predict motion nor the appearance from the given image material. Possible reasons for the poor performance are those mentioned in the preceding section, i.e., the presence of colors, larger difference in pedestrian sizes and, different to SHTC, the changes of the scene itself. As the CHAD dataset was recorded on a parking lot, the parked cars can potentially change, which is exactly what happens here. Analogous to Figure 5.14 two examples of (real,fake) pairs are depicted in Figure 5.15, illustrating the aforementioned situation. In the left setting, the input frame shows a dark car at the left edge of the frame, while the generated frame shows two cars, a silver and a red one. The same applies to the second example, where the input frame contains a van at the lower end of the frame

and a second car right above the van. Both cars are removed from the frame by the MurzGAN. This highlights another weakness of the approach: Since the MurzGAN is based on the assumption that the scene is static, changes to the scene ensure that the network is not able generate correct data.

Figure 5.15: Analogous to Figure 5.14 two examples from the CHAD dataset are illustrated. The changes in the scene due to the changing presence of certain cars cannot be adequately captured by the approach, i.e., it still generates cars despite their absence.

As the SHTC dataset is much more similar to UCSD dataset, changes in the scene do not occur. The table indicates, that MurzGAN is capable of retrieving the desired information from the data. Furthermore the outcomes underline the findings in [Gol19c], that exclusively relying motion information for this particular scenario achieves the best performance, compared to just using the privacy-sensitive frame data or the combination of both.

For visual examples taken from SHTC, CHAD, and UCSD dataset see Figures A.7 and A.8.

5.4.2.2 SBAD: Custom Approaches and Extensions

This next section takes a look on the BinAE model proposed in [Gol22b] and focuses on various aspects related to the model. As presented in Section 4.4.3.1 the base architecture consists of two AEs, one exclusively supplied with information on normal and the other on anomalous motion patterns. To investigate the influence of different introduced parts of the overall custom model, this section provides results on two publicly available datasets: VFP290K and CHAD.

Table 5.16: Ablation study for the BinAE showing the influence of the different extensions. The experiments were performed on the VFP290K dataset.

Config	AUC-ROC	AUC-PR	EER
AE_n	0.6483 ± 0.0854	0.8390 ± 0.0400	0.3675 ± 0.0735
+ Memory	0.6723 ± 0.0294	0.8539 ± 0.0204	0.3454 ± 0.0209
+ AE_a (incl. MLP)	0.9640 ± 0.0183	0.9892 ± 0.0062	0.1099 ± 0.0346
+ Pseudo Anomalies	0.9412 ± 0.0236	0.9818 ± 0.0085	0.1369 ± 0.0448
DualHeadAE	0.9050 ± 0.0647	0.9635 ± 0.0286	0.1529 ± 0.0641

The first three rows of Tables 5.16 and 5.17 show the influence of the different components of the BinAE evaluated on both, the VFP290K and CHAD dataset. First and foremost it should be stated, that the overall performance of the BinAE and its various components and extensions differs strongly between the two regarded datasets. VFP290K seems to be less challenging as the results in Table 5.16 show, which comes mainly from the low diversity in the anomalous and normal movement patterns. Even the very basic AE_n that just gets normal pose sequences to train on achieves high scores on distinguishing falling people or such laying on the ground as the AUC-PR shows. This means that the encoder is already capable of learning a suited embedding containing enough information for the decoder to reconstruct the initial sequence. Notably, other than reported in [Gol22b], extending the AE_n with a memory unit does not result in a significant performance improvement, yet it improves the scores and reduces the error on average, as well as reducing the variance on these results. Completing the BinAE architecture by adding the second Autoencoder AE_a including the final scoring network yields a significantly better result over all calculated metrics. This is a comprehensible observation, as the first two rows are trained in an unsupervised manner, meaning that all of the information the model can use to rely its decision onto is based on examples from the normal behavior cases. Extending the level of information by supplying the model with examples of opposite cases, i.e., such that are known to be anomalous, or belong to the class of relevant behavior as described in Section 4.4.1 helps to distinguish both types of behavior from each other. Although this is not the typical assumption that is made for classical anomaly detection tasks, namely that there is no knowledge about what to expect as an anomaly, it

Table 5.17: Analogous to Table 5.16, an ablation study for the BinAE is showing the influence of the different extensions. The experiments were performed on the CHAD dataset.

Config	AUC-ROC	AUC-PR	EER
AE_n	0.5367 ± 0.0170	0.1700 ± 0.0107	0.4782 ± 0.0090
+ Memory	0.5266 ± 0.0125	0.1580 ± 0.0051	0.4792 ± 0.0112
+ AE_a (incl. MLP)	0.5757 ± 0.0218	0.2194 ± 0.0241	0.4563 ± 0.0170
+ Pseudo Anomalies	0.6040 ± 0.0286	0.2475 ± 0.0575	0.4325 ± 0.0195
DualHeadAE	0.6490 ± 0.0108	0.3716 ± 0.0135	0.4032 ± 0.0081

is completely reasonable to use such information in addition to just relying decisions on those cases that aim on representing normality.

Analogous to Table 5.16, the same experiments were conducted on the CHAD dataset, with its results being reported in Table 5.17. Apparently, the overall range of scores is much lower than observed on the VFP290K dataset, which signifies the increased difficulty and complexity of the CHAD dataset compared to the former. This comes mainly from the way the anomalous samples are composed, in particular the wider range of activities that are seen as anomalous compared to the VFP290K. As described in Section 5.3.3.1, CHAD consists of a higher variety of anomalous behavior patterns, hence the model is expected to recognize a wider range of motion patterns being anomalous or salient. Other than VFP290K the difference between these behaviors is much more subtle, making it harder to distinguish between normal and anomalous samples. Generally speaking the observations that are made on the CHAD are similar to those on the VFP290K dataset. As the scenes in CHAD are much more natural, i.e., with external influences like cars and other objects that can occlude pedestrians at least partially, the utilized pose estimator has to cope with more challenges. This results in specific cases, where the HPE algorithm generates incorrect pose estimates, which will certainly have an effect on the behavior analysis. Notably this is not just an observation that can be made with the chosen HRNet variant but is a fundamental issue for any HPE method.

The next extension to the training that was made is the loss function used during training. Typically, the scenario for the supervised anomaly detection task is shaped by a strong class imbalance where the normal class with way

Table 5.18: Comparison of the the shrinkage loss and the mean squared error loss for training the DualHeadAE on both CHAD and VFP290K dataset.

Dataset	Approach	AUC-ROC	AUC-PR	EER
CHAD	Shrinkage Loss	0.6490 ± 0.0108	0.3716 ± 0.0135	0.4032 ± 0.0081
	Mean Squared Error	0.6518 ± 0.0137	0.3844 ± 0.0181	0.4014 ± 0.0114
VFP290K	Shrinkage Loss	0.9050 ± 0.0647	0.9635 ± 0.0286	0.1529 ± 0.0641
	Mean Squared Error	0.9293 ± 0.0511	0.9754 ± 0.0207	0.1350 ± 0.0569

more samples. During training this way it can happen, that the model concentrates assigning everything to the normal class. To address this issue, the shrinkage loss was introduced. However, as the results displayed in Table 5.18 indicate, on average the overall performance seems to decrease when training rather than improving the performance. Yet it has to be stated, that these are not significant results as the variance of the results shows. This observation can be made on both regarded datasets.

The next group of experiments addresses the topic of privacy-friendliness with respect to the training aspects mainly by adding synthetic samples to the training. To do so, two different approaches were presented, the first one working on existing real-world samples that generates pseudo-anomalies by altering their pose configurations (cf. Section 4.4.3.4) and the second by adding realistic synthetic (cf. Section 4.2.1) samples to the training process. These experiments were performed in particular with the DualHeadAE on the CHAD dataset. The corresponding obtained results are presented in Table 5.19. For both scenarios the fundamental assumption is that such samples help the model to distinguish normal samples from anomalous better by focusing more an the essentials of the normal movements. As the results indicate, the pseudo-anomalies, which are indicated with *pseud.*, do not improve the performance of the DualHeadAE when included in the training. Especially when looking at the AUC-PR shows an negative impact on the training, as the score is significantly smaller. A similar observation can be made using the BinAE on the VFP290K dataset, which decreases in performance when enhanced with pseudo-anomalies as Table 5.16 reveals. The generated pseudo-anomalies are too far from how anomalies that are represented in VFP290K

Table 5.19: Performance impact of the enriched training process. In addition to the existing dataset, synthetic samples obtained from the MixAMoR dataset are used to support the training of the DualHeadAE. The experiments were performed on the CHAD dataset.

Config	AUC-ROC	AUC-PR	EER
DualHeadAE	0.6490 ± 0.0108	0.3716 ± 0.0135	0.4032 ± 0.0081
DualHeadAE + *pseud.*	0.6202 ± 0.0558	0.2621 ± 0.0326	0.4193 ± 0.0438
DualHeadAE + *syn.*	0.6407 ± 0.0112	0.3677 ± 0.0142	0.4122 ± 0.0083
DualHeadAE/NF@*syn.*	0.6914 ± 0.0581	0.4029 ± 0.0464	0.3514 ± 0.0483

look like, which confuses the model rather than providing beneficial information. The permutations that are generated can yield samples that are noisier, seem to be less disturbing for the CHAD dataset. Since the DualHeadAE uses a shared encoder, it seems that the model is not able to generate suitable encodings when adding the pseudo-anomalies to the training. The BinAE at the same time benefits from these pseudo-anomalies with respect to the results reported on the CHAD dataset. Here, both branches are clearly separated, hence each branch can focus on learning an own embedding, without interfering with the other branch.

The MixAMoR dataset on the other hand consists of realistic samples that show a particular similarity to the anomalous samples in CHAD. When finetuning the DualHeadAE on these samples, a slight decrease in performance can be observed as Table 5.19 indicates. Here, the utilization of MixAMoR is indicated by *syn.* in the last two rows of the table. DualHeadAE/NF@*syn.* corresponds to an extended version of the DualHeadAE, which was trained on the CHAD data and combined with an instance of the STG-NF that was trained on the MixAMoR samples. Although it looks like a strong improvement at first glance, the variance in the obtained results indicates that the difference might not be significant. This comes in particular from the STG-NF that yields strongly varying results on the CHAD, which indicates that the actual samples the model gets to see have a strong influence on its performance.

5.4.2.3 Comparison with State-of-the-Art

Next, the final versions of BinAE and DualHeadAE are evaluated on the previously introduced datasets and compared with current state-of-the-art approaches presented in Section 3.4. Note that for the following tables, the results for the GEPC and MPED-RNN are taken from literature and provided for reverence, whereas the results for the STG-NF and MoPRL were obtained from own experiments. Since the results for the first two approaches were taken from literature, they are not really comparable to the other results shown in the table. However, as the procedures are similar, they are still added for reference.

The first comparison of the different approaches is based on SHTC-HR dataset. As Table 5.20 indicates, all regarded methods perform reasonable well with the STG-NF providing the best performance with respect to all computed metrics. Meanwhile, the DualHeadAE shows by far the worst performance of all methods, which is even inferior to the MurzGAN, which however shows a strong variance in the achieved results. This can be explained as mentioned earlier in this section by the tricky training of GANs and its struggle to converge during training. On the other hand, since the DualHeadAE is a supervised approach, it expects samples for the anomaly class. In order to provide such samples on the SHTC-HR dataset, the DualHeadAE uses pseudo anomalies as only examples for the positive class. An explanation to this result is, that these pseudo anomalies have nothing in common with actual anomalous samples, which corresponds to the results and observation made on the VFP290K dataset. In this particular scenario the model does not benefit from using these samples and does not improve its performance. Furthermore these samples seem to suppress the model from learning anything relevant from even the real normal samples, irritating the encoder part which is shared between the two heads.

The next dataset on which the performance of the different methods are compared is the CHAD dataset. As mentioned earlier it is the most challenging dataset for the SBAD evaluation, being at the same time the most comparable scenario to a real-world setting. The look on the results reveals the espe-

Table 5.20: Comparison of different methods for SBAD and VAD on the SHTC-HR dataset. Since the DualHeadAE is a supervised approach that expect samples for both classes, it is trained using pseudo anomalies as presented in Section 4.4.3.4. Furthermore the results for GEPC and MPED-RNN are those reported in [Gol22b], which is why only a single AUC-ROC value is provided and none for the other two evaluation metrics.

Method	AUC-ROC	AUC-PR	EER
GEPC [Mor19]	0.6630	—	—
MPED-RNN [Mar20]	0.7703	—	—
STG-NF [Hir23]	0.8342 ± 0.0016	0.8565 ± 0.0018	0.2534 ± 0.0035
MoPRL [Yu24]	0.8126 ± 0.0029	0.7662 ± 0.0039	0.2592 ± 0.0038
DualHeadAE	0.4112 ± 0.0580	0.4973 ± 0.0292	0.5026 ± 0.0285
MurzGAN [Gol19c]	0.7055 ± 0.0856	0.6185 ± 0.2111	0.3404 ± 0.0679

Table 5.21: Comparison of different methods for SBAD and VAD on the CHAD dataset. GEPC and MPED-RNN do not report results on CHAD and are therefore excluded from the table.

Method	AUC-ROC	AUC-PR	EER
STG-NF [Hir23]	0.6565 ± 0.0060	0.6962 ± 0.0047	0.3923 ± 0.0049
MoPRL [Yu24]	0.6132 ± 0.0189	0.5523 ± 0.0308	0.4246 ± 0.0145
BinAE [Gol22b]	0.5757 ± 0.0218	0.2194 ± 0.0241	0.4563 ± 0.0170
DualHeadAE	0.6490 ± 0.0108	0.3716 ± 0.0135	0.4032 ± 0.0081
MurzGAN [Gol19c]	0.5194 ± 0.0181	0.4995 ± 0.0918	0.4561 ± 0.0181

cially poor performance of the MurzGAN. As mentioned earlier in this section, MurzGAN suffers from a hard and tricky training due to the utilization of a GANs, which results in many different performance results indicated by the high variance. What makes the approach even less favorable is the aforementioned view dependency, which makes the approach usable for static setups, yet needing to be trained on the specific view. Furthermore, the results show that the two state-of-the-art approaches STG-NF and MoPRL perform comparable to the DualHeadAE underlining the competitiveness of the own approach. Again the STG-NF shows very good performance in differentiating between the distribution of normal an anomalous patterns, as well as the motion prior used by the MoPRL, which shows its effectiveness, as it is focused on rating the movement of pedestrians, i.e., typical anomalies that are represented in CHAD.

The final experiments are conducted on the VFP290K dataset, which focuses on falling or laying people. Comparing all of the earlier mentioned

Table 5.22: Comparison of different methods for SBAD and VAD on the VFP290K dataset. The results for GEPC and MPED-RNN are those reported in [Gol22b]. Due to the dynamic character of the datasets, no experiments were performed with the MurzGAN.

Method	AUC-ROC	AUC-PR	EER
GEPC [Mor19]	0.6403	—	—
MPED-RNN [Mar20]	0.6934	—	—
STG-NF [Hir23]	0.7317 ± 0.0421	0.5497 ± 0.0483	0.3372 ± 0.0381
MoPRL [Yu24]	0.5076 ± 0.0791	0.6507 ± 0.0514	0.4965 ± 0.0664
BinAE [Gol22b]	0.9640 ± 0.0183	0.9892 ± 0.0062	0.1099 ± 0.0346
DualHeadAE	0.9293 ± 0.0511	0.9754 ± 0.0207	0.1350 ± 0.0569

approaches on the VFP290K dataset reveals that BinAE and DualHeadAE both achieve by far the best performance in the given setup, whereas the other approaches struggle to capture the necessary details in order to detect the anomalies defined by the VFP290K dataset. As explained earlier, both approaches benefit from the information about the anomalous or relevant behavior patterns. All other methods were trained only with the normal samples, despite the STG-NF, which can be trained in a supervised manner as well. GEPC and MPED-RNN show, that even without positive samples a reasonable performance can be achieved on the VFP290K dataset. This may be even due to the clearly distinguishable normal and anomaly class, which can be easily separated by the approaches. Even the basic AE_n part of the DualHeadAE, which corresponds to the GEPC but without the Dirichlet mixture model for the cluster association, achieves an already good performance. Interestingly this extension of the GEPC does not really seem to provide an advantage on the VFP290K dataset over the AE itself. The worst performance is achieved by the MoPRL, which stumbles across its inherent assumption: The method assumes that the anomalous samples show a certain movement pattern that is based by a generally increased motion speed captured in the pose sequences and modeled in the motion prior. This still works for certain cases, like for example people falling after they have been running, but not such that are already lying or are collapsed. Note that MurzGAN is

not evaluated on the VFP290K dataset, since it is trained per view which do not exist for this dataset since they are recorded from an hand-held camera. Furthermore, analogously to the evaluation on SHTC-HR, the results of the GEPC and MPED-RNN are taken from literature.

5.4.2.4 Qualitative Results

Lastly, this section takes a qualitative look on two datasets, the introduced HeR19 and CaWa18 dataset using the final DualHeadAE model that was trained using the CHAD dataset, since it is the closes fit to the both scenarios. Both datasets differ to a large extent: While the former shows in total only two humans in a scene, where the male person runs towards the female person both encountering each other, the latter shows a crowd with many more people in a single scene.

HeR19 is taken into consideration to have a look on how the DualHeadAE handles subtle changes in behavior. The first scene of the dataset shows the man attacking the woman after he reached her, while the second scene emerges to be a surprise resulting in an exaggerated hug. Figure 5.16 shows the scoring output of the DualHeadAE for the two sequences of the HeR19 dataset. As mentioned above both sequences are very similar to each other showing the same scene. The upper plot depicts how the frame level score output of the DualHeadAE evolves over time on the aggressive scene, while the lower plot does so for the normal scene. Right in the beginning, in both scenes of the HeR19 dataset starts already with a score around 0.75 to 0.86. The reason for this is the person at the top edge that is waiting to enter the scene. The HRNetW48 struggles to estimate a robust pose, resulting in a hard to handle situation for the DualHeadAE. As the man runs towards the center of the scene, the score goes slightly down until the actual fight starts. A look on the lower curve, i.e., the results for the second scene, shows that there is also a peak at the point both persons encounter each other, however still staying at the lower end of the score, indicating that the DualHeadAE

In an analogous matter, Figure 5.17 depicts the score over time on the CaWa18. For the first 100 frames there are two groups of pedestrians overlapping for

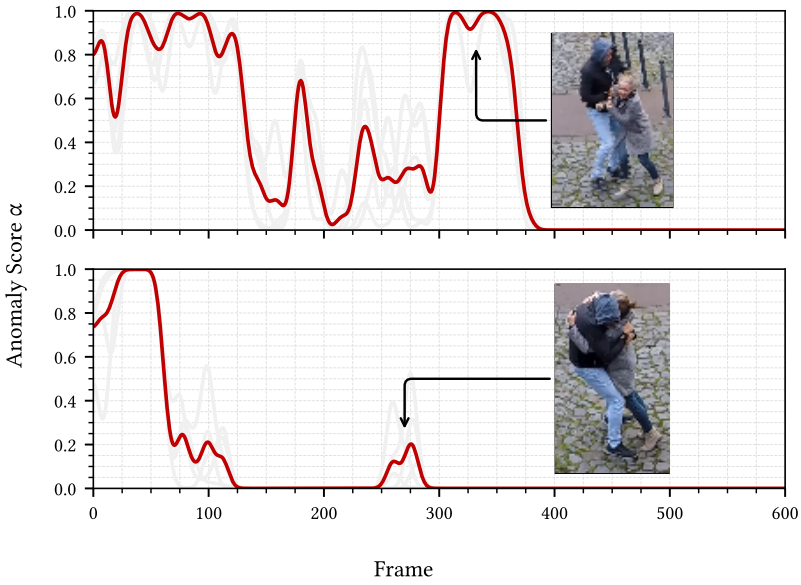


Figure 5.16: DualHeadAE scores over time on both sequences of HeR19. Higher scores indicate anomalous behavior. Apparently the score is high for both scenes right at the beginning coming from the pose estimator not generating a robust and correct pose for the man at the top of the frames.

the exact timespan. After these large groups separate, the scores goes down slightly, still dominated by the many faulty estimations of the pedestrians poses. At around frame 210 to 230 a boy runs in the scene, which is rated with a high score by the DualHeadAE resulting in a slightly increased total score. As already stated for the experiments with regard to the HPE task in Section 5.4.1.2, the dataset is very challenging for the particular task. Consequently, an algorithm like the introduced DualHeadAE that solely relies on the pose estimation has to cope with all problems of the utilized pose estimator. These examples reveal how challenging very crowded situations are for purely pose based analysis. Not only the level of crowdedness plays an important role, but the resolution of the regarded pedestrians. Therefore it is eminently important to have a robust pose estimator. However, as shown

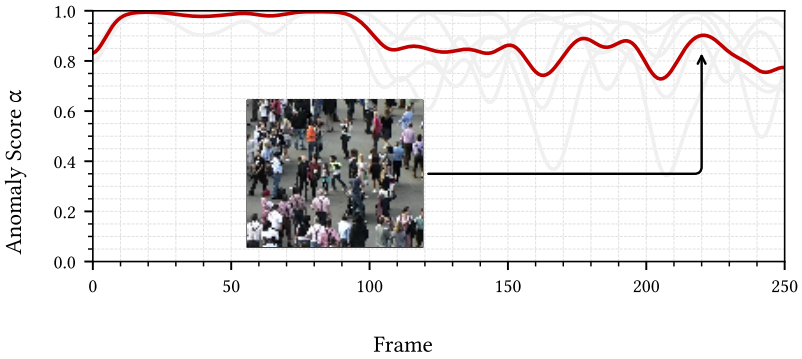


Figure 5.17: DualHeadAE scores over time on the CaWa18. Higher scores indicate anomalous behavior. Apparently the score stays high over the whole scene, which comes mainly from the overwhelmed HRNetW48 which has to estimate robust poses in very crowded situations.

Figure 5.18: Exemplary frames from HeR19 and CaWa18. The left and middle frame show the subtle difference, where one situation (left) is a real attack and the other is just a hug. In contrast the CaWa18 sample shows a situation where the DualHeadAE or any SBA approach has to cope with.

in Section 5.4.1.2, HPE methods in general are struggling with scenarios as the one given by the CaWa18 dataset or other situations where people are overlapping as shown in Figure 5.18. With respect to the privacy-friendly behavior analysis, a better choice might be using a holistic approach rather than a skeleton-based one.

To summarize, the SBAD task is still a very challenging one, however as the results show these kind methods are already suited for certain kind of situations that allow for good separation of people. As the results show, the own approaches are competitive with other state-of-the-art approaches, yet showing still room for improvement. However, especially the attempts to improve the

performance using synthetic augmentation of the training overall did not improve the performance of the DualHeadAE. Further improvement may come from including more effort into the design of the synthetic data with respect to the overall perspective and choice of activities. With respect to the model itself it might be advisable to include concepts like the motion prior [Yu24] or other focused ideas like the normalizing flows [Hir23] as orthogonal concepts that possibly allow in combination with the base DualHeadAE to increase the performance even more, especially for the still challenging cases. As mentioned multiple times, the pose estimation for pedestrians can fail in particularly challenging situations, resulting in unavailable or defect poses, which will have an influence on the classification. However, exactly these cases might be a good choice to use, as the failure of an pose estimator may be a sound indicator that something relevant is going on.

6 Conclusion

This chapter finalizes the dissertation, where Section 6.1 summarizes the contributions and results of this work and reviews them with respect to the overall aim on privacy-friendliness. Section 6.2 gives an outlook on open issues and possibilities to extend the research on privacy-friendliness in human behavior recognition for surveillance scenarios.

6.1 Summary

6.1.1 Contributions and Results

How can we implement privacy-friendly human behavior analysis into a video surveillance world that is defined by privacy concerns as soon as human related video or image material is regarded? This is a question that has shaped and influenced the research conducted in this dissertation, examining various stages of data-driven machine learning approaches as illustrated in Figure 6.1. The central idea that was adapted to achieve this goal is the utilization of human pose information, representing pedestrians as skeletons, which carry as much information as necessary to analyze the behavior of pedestrians.

Although there exists a wide variety of algorithms and machine learning models for the estimation of human poses, their application to urban video surveillance scenarios is particularly challenging as shown in Section 5.4.1 and is rarely regarded by researchers. To address this specific setting, the need for suited and appropriate data arose which had to be handled in a way that is privacy-friendly. For this purpose a synthetic dataset was presented that has

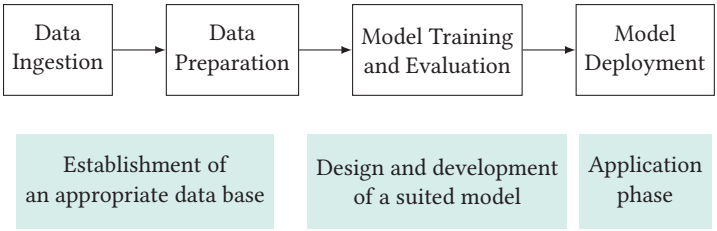


Figure 6.1: Illustration of a typical chain of steps, beginning with the data related operations, the design and training of machine learning models and the final application phase.

been designed to fit the described scenario, i.e., typical urban video surveillance settings, and focuses on providing many pedestrians at the same time with mutual occlusions. Along with this particular dataset a new set of metrics was presented, that aims to describe a given situation with respect to its crowdedness. The so-called Graph Crowd Index generalizes the Crowd Index [Li19] and addresses certain of its drawbacks for the particular scenarios that are typical for video surveillance setups like the elevated view.

As such synthetic data typically shows a certain domain gap towards real-world applications, this dataset was furthermore subject of a domain adaptation process for which a Cycle-GAN-based model was developed that is able to adapt the synthetic data to a desired target domain. The experiments showed, that a carefully designed target domain representation can be beneficial to the overall performance of a HPE approach if trained on such. At the same time, the choice of target domain dataset is crucial, as the DA process may shift the source data still leaving a domain gap. Since the results on HGH18 dataset show, that way it is even possible to outperform recognition performance compared to the training on publicly available datasets. But there is still room for improvement. Particularly not only selecting the appropriate target domain data is crucial, but also the design of the source domain data. The more is known about the appearance of the application domain, the better a suitable source domain dataset can be designed and recorded. As the results presented in Section 5.4.1 showed, the basic appearance of the source domain data is as important as the choice of the right target domain data. Altogether,

generalization in such scenarios is still an issue that has to be addressed in future work.

Last but not least, the actual behavior analysis was addressed with two different kinds of methods, a *holistic* and a *human-centered* approach which were both compared with other state-of-the-art methods, both showing their strengths and weaknesses. While the *holistic* approach suffers from various drawbacks, like difficulties in training due to the GAN-based approach, the *human-centered* method showed competitive performance as presented in Section 5.4.2. In particular, the MurzGAN showed to work well for easier and mostly academic scenarios as defined by datasets like UCSD or SHTC, whereas it was not able to capture the relevant information for scenarios with more subtle behaviors and motion patterns like CHAD, or it was simply not applicable due to the dynamic camera setups, as it is the case for the VFP290K dataset. For the SBAD scenario the experiments showed that it is still a very challenging taks, however as the results show these kind methods are already suited for certain kind of situations that allow for good separation of people. As the results further indicate, the own approach DualHeadAE is competitive with other state-of-the-art approaches, yet showing still room for improvement. In particular it might be advisable to include concepts like the motion prior [Yu24] or other focused ideas like the normalizing flows [Hir23] that possibly allow in combination to increase the performance even more, especially for the currently still challenging cases. Essentially, the pose estimation for pedestrians can fail in particularly challenging situations, resulting in unavailable or defect poses, which will have an influence on the final behavior classification. However, exactly these cases might be a good choice to include into the training, as the failure of an pose estimator may be a sound indicator that something relevant is going on.

Comparing the skeleton-based approaches with the state of the art turned out to be especially challenging, due to the lack of a unified way of benchmarking the various models. This is the reason why such a unified evaluation pipeline was introduced that brings these state-of-the-art methods together providing a benchmark which can be used to assess and compare the performance of arbitrary Skeleton-based Behavior Analysis methods. In particular this pipeline

was applied to compare a selection of state-of-the-art approaches as done in Section 5.4.2. Following the provided benchmark, the results show that taking privacy into consideration when developing models that solely rely on pose information without loss of accuracy is a challenging task, leaving still room for improvement. Especially since following such motivation improves the acceptance of (smart) video surveillance within the society [Gol22a], this is a strong reason that underlines the application of Skeleton-based Behavior Analysis methods particularly for authorities.

In summary this thesis contributes to the research on privacy-aware methodologies for the application field of (smart) video surveillance by addressing various aspects of the development pipeline. The provided behavior analysis pipeline in particular is already suited for real-world application but leaves still room for improvement. However, in its current form it can be already used as a semi-automatic pre-filtering step that helps authorities to process incoming video data. That way the developed methods can be forwarded in future into a more focused classification model, enabling even better notifications that might carry information about the actual activity that was performed. The overall result of this research is just a snapshot showing, that it is already possible to preserve data privacy in modern computer vision system, while achieving reasonable results. However the application of such systems is strongly regulated by the GDPR and local legal rules in different countries, like the police law in Baden-Wuerttemberg in Germany [Lan20].

6.1.2 Review on Privacy-Friendliness

To finalize the thesis, this section takes conclusive look on the privacy-friendliness as leading point for the research at hand. Despite the motivation of using abstract pedestrian representation in terms of 2-dimensional skeletons, there still remain questionable aspects for this choice of feature. Even though skeletons achieve to abstract people from most external influences, like cues on their ethnicity and political orientation, certain other aspects remain recognizable. First and foremost, various studies have proved that re-identifying people is possible based on gait analysis [Elh20, Zha22, Li23].

This may be an issue for future applications of skeleton-based systems as they might lose the eligibility of being classified as privacy-friendly, which would be an even bigger issue, at the latest when it turns out that such information facilitates an identification of people based on their skeletal representations. Furthermore, since the skeletons represent the physical properties of the regarded human being, possibly allowing to recognize whether the person is a child or has a severe physical disability like missing extremities or because the person is sitting in a wheel chair.

Genuinely speaking the behavior analysis pipeline is far from perfect with respect to the privacy topic. At this point, the temporal tracking of individuals is performed based on RGB features, however this can be possibly replaced by other concepts, like classical Bayesian filtering approaches as presented in [Gol21]. Even here it could be possible to use synthetic data to train the inherent motion model represented by the used recurrent neural network, which would allow for reducing the amount of influence and need for RGB data within the behavior analysis pipeline. Furthermore, the attempts to improve the performance by adding synthetic samples showed no real improvement, which calls either for even more effort in generating appropriate samples or for real-world data, where the latter again would be to consider as rather critical with respect to the privacy aspect.

6.2 Outlook

As motivated in the beginning of this thesis, the overall aim is the addressing of privacy concerns in various situations throughout the development of an ideally privacy-friendly behavior analysis pipeline. However, due to the complexity of this task and the variety of aspects that exhibit the potential of bringing risk to the violation of privacy There are still various issues open that have to be addressed in future work. With respect to different topics of this thesis, the following Sections 6.2.1 to 6.2.3 give an outlook on open and future tasks.

6.2.1 Human Feature Representation

With respect to privacy-friendliness in human representations, there is no doubt that using just 2-dimensional flat information as done throughout this thesis is a rather disadvantageous choice and results in certain ambiguities. However, as explained in the introduction, the choice for 2-dimensional skeletons is mainly limited by the current hard- and software systems available and used by the authorities. As presented in Section 3.1.1, there is already research going beyond the established 2-dimensional skeletons, which presents various approaches and provides a variety of datasets that aim on using additional modalities, e.g., in closer setups like driver monitoring in cars. These offer the potential to improve the extraction of human poses or other equivalent features, by using further active or passive sensor systems. However, bringing such technologies closer to application is not just solely an algorithmic issue, but expects companies to develop suited hardware systems that are capable of providing these additional modalities, and obviously authorities to adapt such kind of technologies. Extracting 3-dimensional poses from RGB data, is rather challenging and as the literature on HPE for surveillance and security applications shows, a niche topic. Further research on 3-dimensional pose estimation and the usage of sensors that are able to provide 3-dimensional information about the observed scene is necessary, especially when aiming to convince authorities and companies to focus on adapting such methods. There is no doubt, that using the additional dimension increases robustness of systems and algorithms that aim on interpreting and classifying behavior of pedestrians. The crucial point however is the need for further research has to go beyond solely relying on image data, and adapt further modalities, especially when provided by leading companies[1]. This research should be ideally accompanied by privacy-related studies that further investigate whether these technologies and the resulting algorithms really are as privacy-friendly as they seem in the first place.

[1] AXIS Q1656-DLE uses Radar as additional modality to image data.

6.2.2 Generation of Training Data

As shown in this thesis, generative approaches have the potential in improving the performance of HPE approaches trained on synthetically generated data. GANs have dominated the last decade and have proved their suitability for various tasks, despite the style transfer or domain adaptation as regarded in Section 4.3, there are even further applications of such systems, like e.g., motion transfer. The task of motion transfer goes beyond just altering the style of an image to achieve a better fit to the target domain. While for style transfer it is assumed to have an entirely unlabeled target domain, motion transfer expects the target domain to contain corresponding information, so that these can be changed accordingly. For motion transfer the approaches aim on learning a mapping between a certain colored skeletal representation and the appearance of a particular person. The persons appearance is then transferred into new setups by mapping the appearance on another skeletal representation in a new setting defined by the target domain. By this point, research has been primarily conducted on single person or few people at the same time with each being represented in high resolution video material. As Figure 6.2 shows, adapting this approach for surveillance setups is even more challenging and is underlined by the research presented by Kong [Kon21], who investigated certain of these challenges. Today, seeing the prominent success of foundation models in the field of generative AI might entail further progress for both of these scenarios. Currently, there is only few research on the application of such, especially with respect to the field of video surveillance footage. An appropriate examination of how to use these kinds of models to generate training data based on a synthetic reference dataset will be subject to further research following this dissertation. Despite the algorithmic choice, there is also the design process of the synthetic reference dataset. Although the SyMPose dataset as used in this thesis has been designed to suite the basic target domain, even more time could be invested to tailor an even better fitting dataset. This may include using further GTA modifications, e.g., such that already increase the rendering quality, or completely switching to

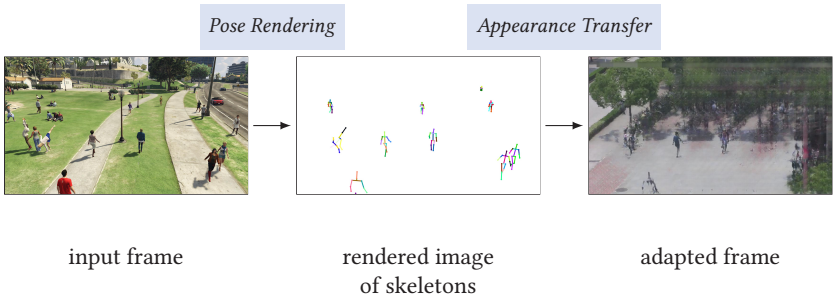


Figure 6.2: Illustration of the conceptual process of motion transfer. First, the skeletons from the source sequence, here SyMPose, are taken and rendered as a single image per frame. These images are created in such way, that each pedestrian shows a certain appearance. The rendering is then used to generate a new frame in the target domain, which is here given by Human in Events dataset [Kon21].

generate data with existing and freely available game engines like Unity[1] or Unreal Engine[2], which however expects the user to understand a much more complex system due to the increased degrees of freedom.

6.2.3 Behavioral Analysis

As mentioned in Section 3.4, skeleton-based methods are becoming ever more popular with respect to video-based behavior analysis. Depending on the actual problem, Skeleton-based Action Recognition and Skeleton-based Anomaly Detection are typical applications, however with the privacy issues in mind, most approaches are still based on RGB data. This may potentially lead the behavioral analysis to base its decision on features that are possibly not directly related to the behavior, like the ethnicity or political orientation of people. While statistically certain correlations can undoubtedly be inferred from the appearance of a person, this kind of preoccupation is no desired virtue for an automatic system. As shown in [Gol22a], the European

[1] https://unity.com/
[2] https://www.unrealengine.com/

population appreciates the efforts of using as few data as possible, and with respect to the AI Act [Eur24] such system can currently be classified as of low risk for the society. These are points that underline the importance of systems as presented in this thesis. SBAD seen as an one-class problem may be used to pre-filter situations before applying much more sophisticated and finegrained classifications as done for SBAR. Although the presented method for SBAD does not use visual cues to rate the relevance or saliency of a given sequence of movements, there are preceding steps that still use pure RGB data, resulting in the same issues as mentioned above. As investigated in [Gol21], one way to reduce the impact of real-world data could be to use synthetic data in combination with an AI-based Bayesian filtering approach [Oka19] to perform the temporal construction of skeleton sequences that can then be analyzed using the presented DualHeadAE approach or any other skeleton-based method as introduced in this thesis.

As stated by Flaborea et al. [Fla23], SBAD should be interpreted as a multi-class, rather than a one-class problem. Anomalous or relevant movement patterns show properties with a wide range of variations, which results in a hard task to distinguish between these intrinsic variations, especially since appropriate data is often not available. Applying generative methods like stable diffusion or normalizing flows to produce pseudo-anomalies may help to overcome the resulting limitations and might even improve behavioral anomaly detection, by interpreting the overall task not just as a one-class problem. Furthermore, the approach presented in this thesis could be improved even further by combining it with zero- or few-shot learning techniques, e.g., CLIP [Rad21], to handle such rare cases, either by real-world samples or by synthetically generated ones as described earlier.

Bibliography

[Ada08] Adam, Amit; Rivlin, Ehud; Shimshoni, Ilan and Reinitz, Daviv: "Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.3 (2008), pp. 555–560. doi: 10.1109/TPAMI.2007.70825 (cit. on p. 124).

[Agg23] Aggarwal, Charu C.: Neural Networks and Deep Learning: A Textbook. Springer International Publishing, 2023. doi: 10.1007/978-3-031-29642-0. url: http://dx.doi.org/10.1007/978-3-031-29642-0 (cit. on pp. 15, 16).

[Aky22] Akyon, Fatih Cagatay; Onur Altinuc, Sinan and Temizel, Alptekin: "Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection". In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2022. doi: 10.1109/icip46576.2022.9897990. url: http://dx.doi.org/10.1109/ICIP46576.2022.9897990 (cit. on p. 132).

[Ala14] Alain, Guillaume and Bengio, Yoshua: "What Regularized Auto-Encoders Learn from the Data-Generating Distribution". In: *Journal of Machine Learning Research* 15.1 (Jan. 2014), pp. 3563–3593. doi: https://dl.acm.org/doi/10.5555/2627435.2750359 (cit. on p. 21).

[Ald22] Aldayri, Amnah and Albattah, Waleed: "Taxonomy of Anomaly Detection Techniques in Crowd Scenes". In: *Sensors* 22.16 (Aug. 2022), p. 6080. doi: 10.3390/s22166080. url: http://dx.doi.org/10.3390/s22166080 (cit. on p. 52).

[Amr23] Amrish; Arya, Shwetank and Kumar, Saurabh: "Convolutional neural network for human crowd analysis: a review". In: *Multimedia Tools and Applications* (Sept. 2023). doi: 10.1007/s11042-023-16841-5. url: http://dx.doi.org/10.1007/s11042-023-16841-5 (cit. on p. 51).

[An21] An, Jaeju; Kim, Jeongho; Lee, Hanbeen; Kim, Jinbeom; Kang, Junhyung; Kim, Minha; Shin, Saebyeol; Kim, Minha; Hong, Donghee and Woo, Simon: "VFP290K: A Large-Scale Benchmark Dataset for Vision-based Fallen Person Detection". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Ed. by Vanschoren, J. and Yeung, S. Vol. 1. 2021. url: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper-round2.pdf (cit. on p. 125).

[And14] Andriluka, Mykhaylo; Pishchulin, Leonid; Gehler, Peter and Schiele, Bernt: "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014. doi: 10.1109/cvpr.2014.471. url: http://dx.doi.org/10.1109/CVPR.2014.471 (cit. on pp. 30, 41, 112).

[And18] Andriluka, Mykhaylo; Iqbal, Umar; Insafutdinov, Eldar; Pishchulin, Leonid; Milan, Anton; Gall, Juergen and Schiele, Bernt: "PoseTrack: A Benchmark for Human Pose Estimation and Tracking". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00542. url: http://dx.doi.org/10.1109/CVPR.2018.00542 (cit. on pp. 30, 42, 43).

[Ang05] Anguelov, Dragomir; Srinivasan, Praveen; Koller, Daphne; Thrun, Sebastian; Rodgers, Jim and Davis, James: "SCAPE: shape completion and animation of people". In: *ACM Transactions on Graphics* 24.3 (July 2005), pp. 408–416. doi: 10.1145/

1073204.1073207. URL: http://dx.doi.org/10.1145/1073204.1073207 (cit. on p. 32).

[Ba16] Ba, Jimmy Lei; Kiros, Jamie Ryan and Hinton, Geoffrey E.: Layer Normalization. 2016. DOI: 10.48550/ARXIV.1607.06450. URL: https://arxiv.org/abs/1607.06450 (cit. on p. 27).

[Ber21a] Bernot, A.: "Transnational state-corporate symbiosis of public security: china's exports of surveillance technologies". In: *International Journal for Crime Justice and Social Democracy* 10 (2 2021). DOI: 10.5204/ijcjsd.1908 (cit. on p. 5).

[Ber21b] Bernot, A.; Trauth-Goik, A. and Trevaskes, S.: "Handling covid-19 with big data in china: increasing 'governance capacity' or 'function creep'?" In: *Australian Journal of International Affairs* 75 (5 2021), pp. 480–486. DOI: 10.1080/10357718.2021.1956430 (cit. on p. 5).

[Bia21] Bian, Yihan and Tang, Xinchen: "Abnormal Detection in Big Data Video with an Improved Autoencoder". In: *Computational Intelligence and Neuroscience* 2021 (Oct. 2021). Ed. by Ding, Bai Yuan, pp. 1–6. DOI: 10.1155/2021/9861533. URL: http://dx.doi.org/10.1155/2021/9861533 (cit. on p. 52).

[Bla19] Blattmann, Andreas: "Multi Person Pose Estimation using Synthetically Generated Data". Master's Thesis. Karlsruhe Institute of Technology, July 2019 (cit. on pp. 62, 63, 65, 79–81, 83–85, 116, 119, 121, 122).

[Boc20] Bochkovskiy, Alexey; Wang, Chien-Yao and Liao, Hong-Yuan Mark: "Yolov4: Optimal speed and accuracy of object detection". In: *arXiv preprint arXiv:2004.10934* (2020) (cit. on p. 127).

[Bra14] Brauer, Jürgen: "Human Pose Estimation with Implicit Shape Models". In: (Apr. 2014). DOI: 10.5445/KSP/1000039083. URL: https://publikationen.bibliothek.kit.edu/1000039083 (cit. on p. 32).

[Cao21] Cao, Zhe; Hidalgo, Gines; Simon, Tomas; Wei, Shih-En and Sheikh, Yaser: "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (Jan. 2021), pp. 172–186. doi: 10.1109/tpami.2019.2929257. url: http://dx.doi.org/10.1109/TPAMI.2019.2929257 (cit. on pp. 34, 36, 234).

[Car20] Carion, Nicolas; Massa, Francisco; Synnaeve, Gabriel; Usunier, Nicolas; Kirillov, Alexander and Zagoruyko, Sergey: End-to-End Object Detection with Transformers. 2020. arXiv: 2005.12872 [cs.CV] (cit. on p. 27).

[Cha20] Chang, Yunpeng; Tu, Zhigang; Xie, Wei and Yuan, Junsong: "Clustering Driven Deep Autoencoder for Video Anomaly Detection". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 329–345. doi: 10.1007/978-3-030-58555-6_20. url: http://dx.doi.org/10.1007/978-3-030-58555-6_20 (cit. on p. 98).

[Che20] Chen, Zhiqian; Chen, Fanglan; Zhang, Lei; Ji, Taoran; Fu, Kaiqun; Zhao, Liang; Chen, Feng and Lu, Chang-Tien: Bridging the Gap between Spatial and Spectral Domains: A Survey on Graph Neural Networks. May 2020 (cit. on pp. 19, 20).

[Cho14] Cho, Kyunghyun; Merrienboer, Bart van; Bahdanau, Dzmitry and Bengio, Yoshua: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 2014. doi: 10.48550/ARXIV.1409.1259. url: https://arxiv.org/abs/1409.1259 (cit. on pp. 22, 24).

[Cho20] Chong, Min Jin and Forsyth, David: Effectively Unbiased FID and Inception Score and where to find them. June 2020. doi: 10.48550/ARXIV.1911.07023. url: https://arxiv.org/abs/1911.07023 (cit. on p. 137).

[Cor16] Cordts, Marius; Omran, Mohamed; Ramos, Sebastian; Rehfeld, Timo; Enzweiler, Markus; Benenson, Rodrigo; Franke, Uwe; Roth, Stefan and Schiele, Bernt: "The Cityscapes Dataset

for Semantic Urban Scene Understanding". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. DOI: 10.1109/cvpr.2016.350. URL: http://dx.doi.org/10.1109/CVPR.2016.350 (cit. on p. 80).

[Cor21] Cormier, Mickael; Röpke, Fabian; Golda, Thomas and Beyerer, Jürgen: "Interactive Labeling for Human Pose Estimation in Surveillance Videos". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021, pp. 1649–1658. DOI: 10.1109/ICCVW54120.2021.00190 (cit. on p. 66).

[Csu17a] Csurka, Gabriela: Domain Adaptation for Visual Applications: A Comprehensive Survey. 2017. DOI: 10.48550/ARXIV.1702.05374. URL: https://arxiv.org/abs/1702.05374 (cit. on p. 48).

[Csu17b] Csurka, Gabriela, ed.: Domain Adaptation in Computer Vision Applications. Springer International Publishing, 2017. DOI: 10.1007/978-3-319-58347-1. URL: https://doi.org/10.1007/978-3-319-58347-1 (cit. on pp. 77, 78).

[Cur20] Curtó, J. D. and Duvall, R.: "Cycle-consistent Generative Adversarial Networks for Neural Style Transfer using data from Chang'E-4". In: *ArXiv* abs/2011.11627 (2020) (cit. on p. 82).

[DAm09] D'Amico, Arnaldo; Di Natale, Corrado; Lo Castro, Fabio; Iarossi, Sergio; Catini, Alexandro and Martinelli, Eugenio: Unexploded Ordnance Detection and Mitigation. Ed. by Byrnes, James. Springer Netherlands, 2009, pp. 21–60. DOI: 10.1007/978-1-4020-9253-4. URL: http://dx.doi.org/10.1007/978-1-4020-9253-4 (cit. on p. 35).

[Dan23] Danesh Pazho, Armin; Alinezhad Noghre, Ghazal; Rahimi Ardabili, Babak; Neff, Christopher and Tabkhi, Hamed: "CHAD: Charlotte Anomaly Dataset". In: *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2023, pp. 50–66. DOI: 10.1007/978-3-031-31435-3_4. URL: http://dx.doi.org/10.1007/978-3-031-31435-3_4 (cit. on pp. 126, 127).

[Den09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai and Fei-Fei, Li: "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009. DOI: 10.1109/cvpr.2009.5206848. URL: http://dx.doi.org/10.1109/CVPR.2009.5206848 (cit. on p. 132).

[Dev18] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton and Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Oct. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805 (cit. on pp. 28, 232).

[Dis18] Dissert, Thomas: "Crowd level Person Pose Estimation". Bachelor's Thesis. Karlsruhe Institute of Technology, Oct. 2018 (cit. on p. 117).

[Dos20] Dosovitskiy, Alexey et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929 (cit. on pp. 27, 28, 148).

[Du23] Du, Yunhao; Zhao, Zhicheng; Song, Yang; Zhao, Yanyun; Su, Fei; Gong, Tao and Meng, Hongying: "StrongSORT: Make DeepSORT Great Again". In: *IEEE Transactions on Multimedia* 25 (Jan. 2023), pp. 8725–8737. DOI: 10.1109/tmm.2023.3240881. URL: http://dx.doi.org/10.1109/TMM.2023.3240881 (cit. on pp. 131, 133, 234).

[Dua21] Duan, Mengmeng; Qiu, Haoyue; Zhang, Zimo and Wu, Yuan: "NTU-DensePose: A New Benchmark for Dense Pose Action Recognition". In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021, pp. 3170–3175. DOI: 10.1109/BigData52589.2021.9671553 (cit. on p. 32).

[Dub22] Dubey, Shradha and Dixit, Manish: "A comprehensive survey on human pose estimation approaches". In: *Multimedia Systems* 29.1 (Aug. 2022), pp. 167–195. DOI: 10.1007/s00530-022-00980-0.

URL: http://dx.doi.org/10.1007/s00530-022-00980-0 (cit. on pp. 31–34).

[Elh20] Elharrouss, Omar; Almaadeed, Noor; Al-Maadeed, Somaya and Bouridane, Ahmed: "Gait recognition for person re-identification". In: *The Journal of Supercomputing* 77.4 (Aug. 2020), pp. 3653–3672. DOI: 10.1007/s11227-020-03409-5. URL: http://dx.doi.org/10.1007/s11227-020-03409-5 (cit. on p. 168).

[Eur24] European Commission: Shaping Europe's digital future - AI Act. [Online; accessed: 22.02.2024]. 2024. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (cit. on pp. 6, 173).

[Fab18] Fabbri, Matteo; Lanzi, Fabio; Calderara, Simone; Palazzi, Andrea; Vezzani, Roberto and Cucchiara, Rita: "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World". In: *European Conference on Computer Vision (ECCV)*. 2018 (cit. on pp. 30, 43, 44, 62).

[Fan23] Fang, Hao-Shu; Li, Jiefeng; Tang, Hongyang; Xu, Chao; Zhu, Haoyi; Xiu, Yuliang; Li, Yong-Lu and Lu, Cewu: "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 2023), pp. 7157–7173. DOI: 10.1109/tpami.2022.3222784. URL: http://dx.doi.org/10.1109/tpami.2022.3222784 (cit. on pp. 34, 232).

[Fen11] Feng, Z.; Li, W. and Varma, J.: "Gaps remain in china's ability to detect emerging infectious diseases despite advances since the onset of sars and avian flu". In: *Health Affairs* 30 (1 2011), pp. 127–135. DOI: 10.1377/hlthaff.2010.0606 (cit. on p. 5).

[Fer18] Fernández, Alberto; García, Salvador; Galar, Mikel; Prati, Ronaldo C; Krawczyk, Bartosz and Herrera, Francisco: Learning from imbalanced data sets. Vol. 10. Springer, 2018 (cit. on p. 114).

[Fla23] Flaborea, Alessandro; Collorone, Luca; D'Amely Di Melendugno, Guido Maria; D'Arrigo, Stefano; Prenkaj, Bardh and Galasso, Fabio: "Multimodal Motion Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023. doi: 10.1109/iccv51070.2023.00947. url: http://dx.doi.org/10.1109/ICCV51070.2023.00947 (cit. on p. 173).

[Foc17] Focus: Bedeutet für Sie mehr Videoüberwachung eher mehr Sicherheit oder eher einen Eingriff in Ihre persönlichen Freiheitsrechte? [Online; accessed: 23.02.2024]. 2017. url: https://de.statista.com/statistik/daten/studie/655303/umfrage/sicherheitsgefuehl-durch-videoueberwachung-in-deutschland/ (cit. on p. 3).

[Fri24] Friedl, Paul and Gasiola, Gustavo Gil: "Examining the EU's Artificial Intelligence Act". In: (Feb. 2024). [Online; accessed: 05.03.2024]. doi: 10.59704/789d6ad759d0a40b. url: http://dx.doi.org/10.59704/789d6ad759d0a40b (cit. on p. 6).

[Fur21] Furst, Michael; Gupta, Shriya T. P.; Schuster, Rene; Wasenmuller, Oliver and Stricker, Didier: "HPERL: 3D Human Pose Estimation from RGB and LiDAR". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Jan. 2021. doi: 10.1109/icpr48806.2021.9412785. url: http://dx.doi.org/10.1109/ICPR48806.2021.9412785 (cit. on p. 37).

[Gao20] Gao, Haoqi and Ogawara, K.: "Generative adversarial network for bidirectional mappings between synthetic and real facial image". In: 11519 (2020), 115190J - 115190J–10. doi: 10.1117/12.2572909 (cit. on p. 82).

[Gat15] Gatys, Leon A.; Ecker, Alexander S. and Bethge, Matthias: A Neural Algorithm of Artistic Style. 2015. doi: 10.48550/ARXIV.1508.06576. url: https://arxiv.org/abs/1508.06576 (cit. on p. 81).

[Gen23] Geng, Jiaqi; Huang, Dong and De la Torre, Fernando: DensePose From WiFi. 2023. doi: 10.48550/ARXIV.2301.00250. url: https://arxiv.org/abs/2301.00250 (cit. on p. 38).

[Geo21] Georgescu, Mariana Iuliana; Ionescu, Radu; Khan, Fahad Shahbaz; Popescu, Marius and Shah, Mubarak: "A Background-Agnostic Framework with Adversarial Training for Abnormal Event Detection in Video". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. doi: 10.1109/tpami.2021.3074805. url: http://dx.doi.org/10.1109/tpami.2021.3074805 (cit. on p. 53).

[Ger00] Gers, Felix A.; Schmidhuber, Jürgen A. and Cummins, Fred A.: "Learning to Forget: Continual Prediction with LSTM". In: *Neural Comput.* 12.10 (Oct. 2000), pp. 2451–2471. doi: 10.1162/089976600300015015. url: https://doi.org/10.1162/089976600300015015 (cit. on p. 22).

[Gla19] Glandon, A.; Vidyaratne, L.; Sadeghzadehyazdi, N.; Dhar, Nibir K.; Familoni, Jide O.; Acton, Scott T. and Iftekharuddin, K. M.: "3D Skeleton Estimation and Human Identity Recognition Using Lidar Full Motion Video". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2019. doi: 10.1109/ijcnn.2019.8852370. url: http://dx.doi.org/10.1109/IJCNN.2019.8852370 (cit. on p. 37).

[Gol19a] Golda, Thomas: "Image-based Anomaly Detection within Crowds". In: *Proceedings of the 2018 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer, M. Taphanel.* Vol. 40. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2019, pp. 11–24. doi: 10.5445/IR/1000097082 (cit. on p. 117).

[Gol19b] Golda, Thomas; Kalb, Tobias; Schumann, Arne and Beyerer, Jürgen: "Human Pose Estimation for Real-World Crowded Scenarios". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019, pp. 1–8. doi: 10.1109/AVSS.2019.8909823 (cit. on pp. 45, 117).

[Gol19c] Golda, Thomas; Murzyn, Nils; Qu, Chengchao and Kroschel, Kristian: "What goes around comes around: Cycle-Consistency-based Short-Term Motion Prediction for Anomaly Detection using Generative Adversarial Networks". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019, pp. 1–8. doi: 10.1109/AVSS.2019.8909853 (cit. on pp. 91, 150–152, 158, 234, 247).

[Gol20] Golda, Thomas; Blattmann, Andreas; Metzler, Jürgen and Beyerer, Jürgen: "Image domain adaption of simulated data for human pose estimation". In: *Artificial Intelligence and Machine Learning in Defense Applications II*. Ed. by Dijk, Judith. Vol. 11543. International Society for Optics and Photonics. SPIE, 2020, pp. 112–127. doi: 10.1117/12.2573888. url: https://doi.org/10.1117/12.2573888 (cit. on pp. 62, 81, 83, 84, 119–121, 134, 137–139).

[Gol21] Golda, Thomas: ""Let's get ready to bundle!: Crowd-level Human Keypoint Tracking". In: *Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer; T. Zander*. Vol. 51. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2021, pp. 67–81. doi: 10.5445/IR/1000135196 (cit. on pp. 169, 173).

[Gol22a] Golda, Thomas; Guaia, Deborah and Wagner-Hartl, Verena: "Perception of Risks and Usefulness of Smart Video Surveillance Systems". In: *Applied Sciences* 12.20 (2022). doi: 10.3390/app122010435. url: https://www.mdpi.com/2076-3417/12/20/10435 (cit. on pp. 3, 59, 168, 172).

[Gol22b] Golda, Thomas; Thiemich, Johanna; Cormier, Mickael and Beyerer, Jürgen: "For the Sake of Privacy: Skeleton-Based Salient Behavior Recognition". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 3983–3987. doi:

10.1109/ICIP46576.2022.9897358 (cit. on pp. 102, 104, 126, 152, 153, 158, 159).

[Gon19] Gong, Dong; Liu, Lingqiao; Le, Vuong; Saha, Budhaditya; Mansour, Moussa Reda; Venkatesh, Svetha and Hengel, Anton van den: "Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection". In: *CoRR* abs/1904.0 (2019). arXiv: 1904.02639. URL: http://arxiv.org/abs/1904.02639 (cit. on pp. 21, 99, 101, 102).

[Goo14] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron and Bengio, Yoshua: "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. and Weinberger, K.Q. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (cit. on pp. 24, 25, 92, 235).

[Goo16] Goodfellow, Ian; Bengio, Yoshua and Courville, Aaron: Deep Learning. http://www.deeplearningbook.org. MIT Press, 2016 (cit. on pp. 15, 16, 21, 110).

[Gra13] Graves, Alex: "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013) (cit. on p. 27).

[Gül18] Güler, Riza Alp; Neverova, Natalia and Kokkinos, Iasonas: "DensePose: Dense Human Pose Estimation in the Wild". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7297–7306. DOI: 10.1109/CVPR.2018.00762 (cit. on p. 32).

[Guo22] Guo, Yongping; Chen, Ying; Deng, Jianzhi; Li, Shuiwang and Zhou, Hui: "Identity-Preserved Human Posture Detection in Infrared Thermal Images: A Benchmark". In: *Sensors* 23.1 (Oct. 2022), p. 92. DOI: 10.3390/s23010092. URL: http://dx.doi.org/10.3390/s23010092 (cit. on p. 36).

[Hao20] Hao, Y. et al.: "Construction and application of surveillance and response systems for parasitic diseases in china, led by nipd-ctdr". In: (2020), pp. 349–371. DOI: 10.1016/bs.apar.2020.04.001 (cit. on p. 5).

[He13] He, Haibo and Ma, Yunqian: "Imbalanced learning: foundations, algorithms, and applications". In: (2013) (cit. on pp. 114, 115).

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing and Sun, Jian: "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 27).

[Hir23] Hirschorn, Or and Avidan, Shai: "Normalizing Flows for Human Pose Anomaly Detection". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023. DOI: 10.1109/iccv51070.2023.01246. URL: http://dx.doi.org/10.1109/ICCV51070.2023.01246 (cit. on pp. 57, 158, 159, 163, 167).

[Hoc97] Hochreiter, Sepp and Schmidhuber, Jürgen: "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 22).

[Hoc98] Hochreiter, Sepp: "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02 (1998), pp. 107–116. DOI: 10.1142/S0218488598000094. eprint: https://doi.org/10.1142/S0218488598000094. URL: https://doi.org/10.1142/S0218488598000094 (cit. on p. 22).

[Hof17] Hoffman, Judy; Tzeng, Eric; Park, Taesung; Zhu, Jun-Yan; Isola, Phillip; Saenko, Kate; Efros, Alexei A. and Darrell, Trevor: "CyCADA: Cycle-Consistent Adversarial Domain Adaptation". In: *CoRR* abs/1711.03213 (2017). arXiv: 1711.03213. URL: http://arxiv.org/abs/1711.03213 (cit. on pp. 50, 51).

[Hoh24] Hohloch, David: "Skeleton-based Behavior Analysis". Bachelor's Thesis. Karlsruhe Institute of Technology, Jan. 2024 (cit. on p. 116).

[Hor89] Hornik, Kurt; Stinchcombe, Maxwell and White, Halbert: "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208 (cit. on p. 15).

[Hua21] Huang, J. and Tsai, K.: "Upgrading big brother: local strategic adaptation in china's security industry". In: *Studies in Comparative International Development* 56 (4 2021), pp. 560–587. DOI: 10.1007/s12116-021-09342-9 (cit. on p. 5).

[Ion14] Ionescu, Catalin; Papava, Dragos; Olaru, Vlad and Sminchisescu, Cristian: "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339. DOI: 10.1109/TPAMI.2013.248 (cit. on p. 31).

[Ion19] Ionescu, Radu Tudor; Khan, Fahad Shahbaz; Georgescu, Mariana-Iuliana and Shao, Ling: "Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. DOI: 10.1109/cvpr.2019.00803. URL: http://dx.doi.org/10.1109/CVPR.2019.00803 (cit. on p. 98).

[Isk19] Iskakov, Karim; Burkov, Egor; Lempitsky, Victor and Malkov, Yury: "Learnable Triangulation of Human Pose". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 7717–7726. DOI: 10.1109/ICCV.2019.00781 (cit. on p. 31).

[Iso17] Isola, Phillip; Zhu, Jun-Yan; Zhou, Tinghui and Efros, Alexei A.: "Image-to-Image Translation with Conditional Adversarial Networks". In: *2017 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR). IEEE, July 2017. DOI: 10.1109/cvpr.2017.632. URL: http://dx.doi.org/10.1109/CVPR.2017.632 (cit. on pp. 82, 92, 93, 95, 234).

[Joc22] Jocher, Glenn et al.: ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. Version v7.0. Nov. 2022. DOI: 10.5281/zenodo.7347926. URL: https://doi.org/10.5281/zenodo.7347926 (cit. on pp. 131, 132).

[Joo18] Joo, Hanbyul; Simon, Tomas and Sheikh, Yaser: "Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. DOI: 10.1109/cvpr.2018.00868. URL: http://dx.doi.org/10.1109/CVPR.2018.00868 (cit. on p. 32).

[Jun20] Jung, Im Y.: "A review of privacy-preserving human and human activity recognition". In: *International Journal on Smart Sensing and Intelligent Systems* 13.1 (May 2020), pp. 1–13. DOI: 10.21307/ijssis-2020-008. URL: http://dx.doi.org/10.21307/ijssis-2020-008 (cit. on pp. 8, 9).

[Kal19] Kalb, Tobias: "Optimization of Human Body Pose Estimation for Crowd Applications". Master's Thesis. Karlsruhe Institute of Technology, May 2019 (cit. on pp. 43–45, 110, 117).

[Kan22] Kanu-Asiegbu, Asiegbu Miracle; Vasudevan, Ram and Du, Xiaoxiao: "Leveraging Trajectory Prediction for Pedestrian Video Anomaly Detection". In: (2022). DOI: 10.48550/ARXIV.2207.02279. URL: https://arxiv.org/abs/2207.02279 (cit. on p. 52).

[Kar21] Karras, Tero; Laine, Samuli and Aila, Timo: "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2021), pp. 4217–4228. DOI: 10.1109/TPAMI.2020.2970919 (cit. on p. 24).

[Kat19] Katovich, Kristina: "Artifical Data for Activity Recognition". Bachelor's Thesis. Karlsruhe Institute of Technology, Nov. 2019 (cit. on p. 60).

[Kie21] Kiefer, Benjamin; Ott, David and Zell, Andreas: Leveraging Synthetic Data in Object Detection on Unmanned Aerial Vehicles. 2021. doi: 10.48550/ARXIV.2112.12252. url: https://arxiv.org/abs/2112.12252 (cit. on p. 62).

[Kim23] Kim, Gon Woo; Lee, Sang Won; Son, Ha Young and Choi, Kae Won: "A Study on 3D Human Pose Estimation Using Through-Wall IR-UWB Radar and Transformer". In: *IEEE Access* 11 (2023), pp. 15082–15095. doi: 10.1109/access.2023.3244017. url: http://dx.doi.org/10.1109/ACCESS.2023.3244017 (cit. on p. 39).

[Kin14] Kingma, Diederik P. and Ba, Jimmy: Adam: A Method for Stochastic Optimization. 2014. doi: 10.48550/ARXIV.1412.6980. url: https://arxiv.org/abs/1412.6980 (cit. on p. 122).

[Kon21] Kong, Xiaoyan: "Generative Adversarial Network based Image Domain Adaption for Multi Person Pose Estimation". Master's Thesis. Karlsruhe Institute of Technology, June 2021 (cit. on pp. 171, 172).

[Kos21] Kostka, G.; Steinacker, L. and Meckel, M.: "Between security and convenience: facial recognition technology in the eyes of citizens in china, germany, the united kingdom, and the united states". In: *Public Understanding of Science* 30 (6 2021), pp. 671–690. doi: 10.1177/09636625211001555 (cit. on p. 5).

[Kri12] Krizhevsky, Alex; Sutskever, Ilya and Hinton, Geoffrey E.: "ImageNet classification with deep convolutional neural networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105 (cit. on pp. 93, 231).

[Lan20] Land Baden-Württemberg: § 44 - Offener Einsatz technischer Mittel zur Bild- und Tonaufzeichnung. https://www.landesrecht-bw.de/bsbw/document/jlr-PolGBW2021pP44, accessed: 05.06.2024. Oct. 2020 (cit. on p. 168).

[Lea15] Leal-Taixé, Laura; Milan, Anton; Reid, Ian D.; Roth, Stefan and Schindler, Konrad: "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking". In: *CoRR* abs/1504.01942 (2015). arXiv: 1504.01942. URL: http://arxiv.org/abs/1504.01942 (cit. on pp. 45, 46).

[Lee18] Lee, Sangmin; Kim, Hak Gu and Ro, Yong Man: "STAN: Spatio-Temporal Adversarial Networks for Abnormal Event Detection". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2018. DOI: 10.1109/icassp.2018.8462388. URL: http://dx.doi.org/10.1109/ICASSP.2018.8462388 (cit. on pp. 53, 94).

[Lee23] Lee, Shih-Po; Kini, Niraj Prakash; Peng, Wen-Hsiao; Ma, Ching-Wen and Hwang, Jenq-Neng: "HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023. DOI: 10.1109/wacv56688.2023.00567. URL: http://dx.doi.org/10.1109/WACV56688.2023.00567 (cit. on p. 39).

[Li13] Li, Weixin; Mahadevan, Vijay and Vasconcelos, Nuno: "Anomaly detection and localization in crowded scenes". In: *IEEE transactions on pattern analysis and machine intelligence* 36.1 (2013), pp. 18–32 (cit. on pp. 124, 151).

[Li19] Li, Jiefeng; Wang, Can; Zhu, Hao; Mao, Yihuan; Fang, Hao-Shu and Lu, Cewu: "CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10855–10864. DOI: 10.1109/CVPR.2019.01112 (cit. on pp. 42, 45, 46, 117, 121, 138, 139, 166).

[Li20] Li, Bo; Leroux, Sam and Simoens, Pieter: Decoupled Appearance and Motion Learning for Efficient Anomaly Detection in Surveillance Video. 2020. DOI: 10.48550/ARXIV.2011.05054. URL: https://arxiv.org/abs/2011.05054 (cit. on pp. 51, 52).

[Li23] Li, Weijia; Hou, Saihui; Zhang, Chunjie; Cao, Chunshui; Liu, Xu; Huang, Yongzhen and Zhao, Yao: "An In-Depth Exploration of Person Re-Identification and Gait Recognition in Cloth-Changing Conditions". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. DOI: 10.1109/cvpr52729.2023.01328. URL: http://dx.doi.org/10.1109/CVPR52729.2023.01328 (cit. on p. 168).

[Lia14] Liang, S.; Yang, C.; Zhong, B.; Guo, J.; Li, H.; Carlton, E.; Freeman, M. and Remais, J.: "Surveillance systems for neglected tropical diseases: global lessons from china's evolving schistosomiasis reporting systems, 1949–2014". In: *Emerging Themes in Epidemiology* 11 (1 2014). DOI: 10.1186/1742-7622-11-19 (cit. on p. 5).

[Lin15] Lin, Tsung-Yi; Maire, Michael; Belongie, Serge; Bourdev, Lubomir; Girshick, Ross; Hays, James; Perona, Pietro; Ramanan, Deva; Zitnick, C. Lawrence and Dollár, Piotr: Microsoft COCO: Common Objects in Context. 2015. arXiv: 1405.0312 [cs.CV] (cit. on pp. 30, 40, 127, 132).

[Lin18] Lin, Tsung-Yi; Goyal, Priya; Girshick, Ross; He, Kaiming and Dollár, Piotr: Focal Loss for Dense Object Detection. Feb. 2018. DOI: 10.48550/ARXIV.1708.02002. URL: https://arxiv.org/abs/1708.02002 (cit. on p. 102).

[Liu18] Liu, W.; W. Luo, D. Lian and Gao, S.: "Future Frame Prediction for Anomaly Detection – A New Baseline". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 94, 124).

[Liu20] Liu, Ziyu; Zhang, Hongwen; Chen, Zhenghao; Wang, Zhiyong and Ouyang, Wanli: "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. DOI: 10.1109/cvpr42600.2020.00022. URL: http://dx.doi.org/10.1109/CVPR42600.2020.00022 (cit. on p. 97).

[Liu21a] Liu, Chengming; Fu, Ronghua; Li, Yinghao; Gao, Yufei; Shi, Lei and Li, Weiwei: "A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection". In: *Applied Sciences* 12.1 (Dec. 2021), p. 4. doi: 10.3390/app12010004. url: http://dx.doi.org/10.3390/app12010004 (cit. on p. 56).

[Liu21b] Liu, Ze; Lin, Yutong; Cao, Yue; Hu, Han; Wei, Yixuan; Zhang, Zheng; Lin, Stephen and Guo, Baining: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021. arXiv: 2103.14030 [cs.CV] (cit. on p. 27).

[Liu21c] Liu, Ze; Lin, Yutong; Cao, Yue; Hu, Han; Wei, Yixuan; Zhang, Zheng; Lin, Stephen and Guo, Baining: "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021. doi: 10.1109/iccv48922.2021.00986. url: http://dx.doi.org/10.1109/ICCV48922.2021.00986 (cit. on p. 148).

[Liu22a] Liu, C.: "Who supports expanding surveillance? exploring public opinion of chinese social credit systems". In: *International Sociology* 37 (3 2022), pp. 391–412. doi: 10.1177/02685809221084446 (cit. on p. 5).

[Liu22b] Liu, Zhuang; Mao, Hanzi; Wu, Chao-Yuan; Feichtenhofer, Christoph; Darrell, Trevor and Xie, Saining: A ConvNet for the 2020s. 2022. doi: 10.48550/ARXIV.2201.03545. url: https://arxiv.org/abs/2201.03545 (cit. on p. 148).

[Liu23] Liu, Wenrui; Chang, Hong; Ma, Bingpeng; Shan, Shiguang and Chen, Xilin: Diversity-Measurable Anomaly Detection. 2023. doi: 10.48550/ARXIV.2303.05047. url: https://arxiv.org/abs/2303.05047 (cit. on p. 54).

[Lon19] Longman, R. and Ptucha, Raymond W.: "Embedded Cyclegan For Shape-Agnostic Image-To-Image Translation". In: *2019 IEEE International Conference on Image Processing (ICIP)* (2019), pp. 969–973. doi: 10.1109/ICIP.2019.8803082 (cit. on p. 82).

[Lop15] Loper, Matthew; Mahmood, Naureen; Romero, Javier; Pons-Moll, Gerard and Black, Michael J.: "SMPL: a skinned multi-person linear model". In: *ACM Transactions on Graphics* 34.6 (Oct. 2015), pp. 1–16. doi: 10.1145/2816795.2818013. url: http://dx.doi.org/10.1145/2816795.2818013 (cit. on p. 32).

[Lu13] Lu, Cewu; Shi, Jianping and Jia, Jiaya: "Abnormal Event Detection at 150 FPS in MATLAB". In: *2013 IEEE International Conference on Computer Vision*. IEEE, Oct. 2013. doi: 10.1109/iccv.2013.338. url: http://dx.doi.org/10.1109/ICCV.2013.338 (cit. on p. 124).

[Lu20] Lu, Xiankai; Ma, Chao; Shen, Jianbing; Yang, Xiaokang; Reid, Ian and Yang, Ming-Hsuan: "Deep Object Tracking with Shrinkage Loss". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1. doi: 10.1109/tpami.2020.3041332. url: http://dx.doi.org/10.1109/TPAMI.2020.3041332 (cit. on p. 103).

[Luo21] Luo, Weixin; Liu, Wen and Gao, Shenghua: "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection". In: *Neurocomputing* 444 (July 2021), pp. 332–337. doi: 10.1016/j.neucom.2019.12.148. url: http://dx.doi.org/10.1016/j.neucom.2019.12.148 (cit. on p. 56).

[Lup24] Lupión, Marcos; Polo-Rodríguez, Aurora; Medina-Quero, Javier; Sanjuan, Juan F. and Ortigosa, Pilar M.: "3D Human Pose Estimation from multi-view thermal vision sensors". In: *Information Fusion* 104 (Apr. 2024), p. 102154. doi: 10.1016/j.inffus.2023.102154. url: http://dx.doi.org/10.1016/j.inffus.2023.102154 (cit. on p. 36).

[Maa08] Maaten, Laurens van der and Hinton, Geoffrey: "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. url: http://jmlr.org/papers/v9/vandermaaten08a.html (cit. on p. 86).

[Mao17] Mao, Xudong; Li, Qing; Xie, Haoran; Lau, Raymond Y.K.; Wang, Zhen and Smolley, Stephen Paul: "Least Squares Generative Adversarial Networks". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Aug. 2017. DOI: 10.1109/iccv.2017.304. URL: http://dx.doi.org/10.1109/ICCV.2017.304 (cit. on p. 92).

[Mar17] Martinez, Julieta; Hossain, Rayat; Romero, Javier and Little, James J.: "A Simple Yet Effective Baseline for 3d Human Pose Estimation". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288 (cit. on p. 31).

[Mar20] Markovitz, Amir; Sharir, Gilad; Friedman, Itamar; Zelnik-Manor, Lihi and Avidan, Shai: "Graph Embedded Pose Clustering for Anomaly Detection". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. DOI: 10.1109/cvpr42600.2020.01055. URL: http://dx.doi.org/10.1109/CVPR42600.2020.01055 (cit. on pp. 21, 56, 98, 158, 159).

[Mat21] Mateus, Balduíno César; Mendes, Mateus; Farinha, José Torres; Assis, Rui and Cardoso, António Marques: "Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press". In: *Energies* 14.21 (Oct. 2021), p. 6958. DOI: 10.3390/en14216958. URL: https://doi.org/10.3390/en14216958 (cit. on p. 24).

[Mes18] Mescheder, Lars M.; Geiger, Andreas and Nowozin, Sebastian: "Which Training Methods for GANs do actually Converge?" In: *International Conference on Machine Learning*. 2018 (cit. on p. 26).

[Mor19] Morais, Romero; Le, Vuong; Tran, Truyen; Saha, Budhaditya; Mansour, Moussa and Venkatesh, Svetha: "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. DOI: 10.1109/cvpr.2019.

01227. URL: http://dx.doi.org/10.1109/CVPR.2019.01227 (cit. on pp. 55, 124, 158, 159).

[Nas51] NASH, J.F.: "Non-cooperative Games". In: *Annals of Mathematics* 54.2 (1951), pp. 286–295 (cit. on p. 25).

[Nat10] NATIONAL AERONAUTICS AND SPACE ADMINISTRATION: Introduction to the Electromagnetic Spectrum. [Online; accessed: 23.12.2023]. 2010. URL: http://science.nasa.gov/ems/01_intro (cit. on p. 35).

[Nev19] NEVEROVA, Natalia; THEWLIS, James; GÜLER, Rıza Alp; KOKKINOS, Iasonas and VEDALDI, Andrea: "Slim DensePose: Thrifty Learning From Sparse Annotations and Motion Cues". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10907–10915. DOI: 10.1109/CVPR.2019.01117 (cit. on p. 32).

[Oka19] OKADA, Masashi; TAKENAKA, Shinji and TANIGUCHI, Tadahiro: Multi-person Pose Tracking using Sequential Monte Carlo with Probabilistic Neural Pose Predictor. 2019. arXiv: 1909.07031 [cs.CV] (cit. on p. 173).

[Oza21] OZA, Poojan; SINDAGI, Vishwanath A.; VS, Vibashan and PATEL, Vishal M.: Unsupervised Domain Adaptation of Object Detectors: A Survey. 2021. DOI: 10.48550/ARXIV.2105.13502. URL: https://arxiv.org/abs/2105.13502 (cit. on p. 48).

[Pan10] PAN, Sinno Jialin and YANG, Qiang: "A Survey on Transfer Learning". In: *IEEE Trans. on Knowl. and Data Eng.* 22.10 (Oct. 2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191. URL: http://dx.doi.org/10.1109/TKDE.2009.191 (cit. on p. 78).

[Paz22] PAZHO, Armin Danesh; NOGHRE, Ghazal Alinezhad; ARDABILI, Babak Rahimi; NEFF, Christopher and TABKHI, Hamed: "CHAD: Charlotte Anomaly Dataset". In: (2022). DOI: 10.48550/ARXIV.2212.09258. URL: https://arxiv.org/abs/2212.09258 (cit. on pp. 113, 126, 128).

[Rad21] Radford, Alec et al.: "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina and Zhang, Tong. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 8748–8763. url: https://proceedings.mlr.press/v139/radford21a.html (cit. on p. 173).

[Rae16] Rae, Jack W; Hunt, Jonathan J; Harley, Tim; Danihelka, Ivo; Senior, Andrew; Wayne, Greg; Graves, Alex and Lillicrap, Timothy P: "Scaling Memory-Augmented Neural Networks with Sparse Reads and Writes". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3628–3636 (cit. on p. 99).

[Rak21] Rakhimov, Ruslan; Bogomolov, Emil; Notchenko, Alexandr; Mao, Fung; Artemov, Alexey; Zorin, Denis and Burnaev, Evgeny: "Making DensePose fast and light". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 1868–1876. doi: 10.1109/WACV48630.2021.00191 (cit. on p. 32).

[Rav17a] Ravanbakhsh, Mahdyar; Nabi, Moin; Sangineto, Enver; Marcenaro, Lucio; Regazzoni, Carlo and Sebe, Nicu: "Abnormal event detection in videos using generative adversarial nets". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2017. doi: 10.1109/icip.2017.8296547. url: http://dx.doi.org/10.1109/ICIP.2017.8296547 (cit. on pp. 52, 53, 91, 93, 94).

[Rav17b] Ravanbakhsh, Mahdyar; Sangineto, Enver; Nabi, Moin and Sebe, Nicu: Training Adversarial Discriminators for Cross-channel Abnormal Event Detection in Crowds. 2017. doi: 10.48550/ARXIV.1706.07680. url: https://arxiv.org/abs/1706.07680 (cit. on pp. 52, 53, 94).

[Red16] Redmon, Joseph; Divvala, Santosh; Girshick, Ross and Farhadi, Ali: "You Only Look Once: Unified, Real-Time Object

Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. DOI: 10.1109/cvpr.2016.91. URL: http://dx.doi.org/10.1109/CVPR.2016.91 (cit. on p. 33).

[Rei22] Reiss, Tal and Hoshen, Yedid: Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection. 2022. DOI: 10.48550/ARXIV.2212.00789. URL: https://arxiv.org/abs/2212.00789 (cit. on pp. 52, 54).

[Ris21] Ristea, Nicolae-Catalin; Madan, Neelu; Ionescu, Radu Tudor; Nasrollahi, Kamal; Khan, Fahad Shahbaz; Moeslund, Thomas B. and Shah, Mubarak: Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection. 2021. DOI: 10.48550/ARXIV.2111.09099. URL: https://arxiv.org/abs/2111.09099 (cit. on p. 52).

[Rod20] Rodrigues, Royston; Bhargava, Neha; Velmurugan, Rajbabu and Chaudhuri, Subhasis: "Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2020. DOI: 10.1109/wacv45572.2020.9093633. URL: http://dx.doi.org/10.1109/WACV45572.2020.9093633 (cit. on p. 56).

[Ron15] Ronneberger, Olaf; Fischer, Philipp and Brox, Thomas: "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28. URL: http://dx.doi.org/10.1007/978-3-319-24574-4_28 (cit. on pp. 82, 235).

[Ron17] Ronchi, Matteo Ruggero and Perona, Pietro: "Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation". In: *CoRR* abs/1707.05388 (2017). arXiv: 1707.05388. URL: http://arxiv.org/abs/1707.05388 (cit. on pp. 41, 111–113).

[Sai15] Saito, Takaya and Rehmsmeier, Marc: "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432 (cit. on p. 115).

[Sal16] Salimans, Tim; Goodfellow, Ian J.; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec and Chen, Xi: "Improved Techniques for Training GANs". In: *CoRR* abs/1606.03498 (2016) (cit. on p. 25).

[San18] Sanakoyeu, Artsiom; Kotovenko, Dmytro; Lang, Sabine and Ommer, Björn: A Style-Aware Content Loss for Real-time HD Style Transfer. 2018. doi: 10.48550/ARXIV.1807.10201. url: https://arxiv.org/abs/1807.10201 (cit. on p. 83).

[Ser86] Serra, Jean: "Introduction to mathematical morphology". In: *Computer Vision, Graphics, and Image Processing* 35.3 (Sept. 1986), pp. 283–305. doi: 10.1016/0734-189x(86)90002-2. url: http://dx.doi.org/10.1016/0734-189X(86)90002-2 (cit. on pp. 91, 96).

[Sha22] Sharif, Md. Haidar; Jiao, Lei and Omlin, Christian W.: Deep Crowd Anomaly Detection: State-of-the-Art, Challenges, and Future Research Directions. 2022. doi: 10.48550/ARXIV.2210.13927. url: https://arxiv.org/abs/2210.13927 (cit. on pp. 113, 247).

[She17] Shelhamer, Evan; Long, Jonathan and Darrell, Trevor: "Fully Convolutional Networks for Semantic Segmentation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (Apr. 2017), pp. 640–651. doi: 10.1109/TPAMI.2016.2572683. url: https://doi.org/10.1109/TPAMI.2016.2572683 (cit. on p. 20).

[She21] Sheng, Yiwei: "Asymmetric CycleGAN for Unpaired Image-to-Image Translation Based on Dual Attention Module". In: *2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)* (2021), pp. 726–730. doi: 10.1109/iaecst54258.2021.9695748 (cit. on p. 82).

[Sid00] Sidenbladh, H.; De la Torre, F. and Black, M.J.: "A framework for modeling the appearance of 3D articulated figures". In: *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. AFGR-00. IEEE Comput. Soc, 2000. DOI: 10.1109/afgr.2000.840661. URL: http://dx.doi.org/10.1109/AFGR.2000.840661 (cit. on p. 32).

[Sig09] Sigal, Leonid; Balan, Alexandru O. and Black, Michael J.: "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion". In: *International Journal of Computer Vision* 87.1-2 (Aug. 2009), pp. 4–27. DOI: 10.1007/s11263-009-0273-6. URL: https://doi.org/10.1007/s11263-009-0273-6 (cit. on p. 31).

[Sim14] Simonyan, Karen and Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. DOI: 10.48550/ARXIV.1409.1556. URL: https://arxiv.org/abs/1409.1556 (cit. on pp. 86, 93, 148).

[Sin23] Singh, Rituraj; Saini, Krishanu; Sethi, Anikeit; Tiwari, Aruna; Saurav, Sumeet and Singh, Sanjay: "STemGAN: spatio-temporal generative adversarial network for video anomaly detection". In: *Applied Intelligence* 53.23 (Sept. 2023), pp. 28133–28152. DOI: 10.1007/s10489-023-04940-7. URL: http://dx.doi.org/10.1007/s10489-023-04940-7 (cit. on p. 53).

[Smi23] Smith, Javier; Loncomilla, Patricio and Ruiz-Del-Solar, Javier: "Human Pose Estimation Using Thermal Images". In: *IEEE Access* 11 (2023), pp. 35352–35370. DOI: 10.1109/access.2023.3264714. URL: http://dx.doi.org/10.1109/ACCESS.2023.3264714 (cit. on p. 36).

[Spi20] Spiegel: Wie stehen Sie zu einem Einsatz von automatisierter Gesichtserkennung durch Behörden? [Online; accessed: 13.05.2024]. 2020. URL: https://de.statista.com/statistik/daten/studie/1091820/umfrage/umfrage-zum-einsatz-von-automatisierter-gesichtserkennung-durch-behoerden/ (cit. on p. 3).

[Sta22] Starke, Sebastian; Mason, Ian and Komura, Taku: "DeepPhase: Periodic Autoencoders for Learning Motion Phase Manifolds". In: *ACM Trans. Graph.* 41.4 (Oct. 2022). doi: 10.1145/3528223.3530178. url: https://doi.org/10.1145/3528223.3530178 (cit. on p. 62).

[Sul18] Sultani, Waqas; Chen, Chen and Shah, Mubarak: "Real-World Anomaly Detection in Surveillance Videos". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00678. url: http://dx.doi.org/10.1109/CVPR.2018.00678 (cit. on p. 116).

[Sun17] Sun, Chen; Shrivastava, Abhinav; Singh, Saurabh and Gupta, Abhinav: "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. doi: 10.1109/iccv.2017.97. url: http://dx.doi.org/10.1109/ICCV.2017.97 (cit. on pp. 29, 234).

[Sun19] Sun, Ke; Xiao, Bin; Liu, Dong and Wang, Jingdong: "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 5693–5703. doi: 10.1109/cvpr.2019.00584. url: http://dx.doi.org/10.1109/cvpr.2019.00584 (cit. on pp. 34, 39, 122, 127, 131, 138, 231).

[Tom17] Tome, Denis; Russell, Chris and Agapito, Lourdes: "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5689–5698. doi: 10.1109/CVPR.2017.603 (cit. on p. 31).

[Ton15] Tong, M. et al.: "Infectious diseases, urbanization and climate change: challenges in future china". In: *International Journal of Environmental Research and Public Health* 12 (9 2015), pp. 11025–11036. doi: 10.3390/ijerph120911025 (cit. on p. 5).

[Tou21] Touvron, Hugo; Cord, Matthieu; Douze, Matthijs; Massa, Francisco; Sablayrolles, Alexandre and Jégou, Hervé: Training data-efficient image transformers & distillation through attention. 2021. arXiv: 2012.12877 [cs.CV] (cit. on p. 27).

[Vas23] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz and Polosukhin, Illia: Attention Is All You Need. 2023. arXiv: 1706.03762 [cs.CL] (cit. on p. 27).

[Wan18] Wang, Mei and Deng, Weihong: "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (Oct. 2018), pp. 135–153. doi: 10.1016/j.neucom.2018.05.083. url: http://dx.doi.org/10.1016/j.neucom.2018.05.083 (cit. on p. 48).

[Wan19] Wang, Qi; Gao, Junyu; Lin, Wei and Yuan, Yuan: "Learning from Synthetic Data for Crowd Counting in the Wild". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8198–8207 (cit. on p. 62).

[Wan22] Wang, Guodong; Wang, Yunhong; Qin, Jie; Zhang, Dongming; Bao, Xiuguo and Huang, Di: "Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles". In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 494–511. doi: 10.1007/978-3-031-20080-9_29. url: http://dx.doi.org/10.1007/978-3-031-20080-9_29 (cit. on pp. 52, 54, 104).

[Wei16a] Wei, Shih-En; Ramakrishna, Varun; Kanade, Takeo and Sheikh, Yaser: "Convolutional Pose Machines". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4724–4732. doi: 10.1109/CVPR.2016.511 (cit. on p. 31).

[Wei16b] Weiss, Karl; Khoshgoftaar, Taghi M. and Wang, DingDing: "A survey of transfer learning". In: *Journal of Big Data* 3.1 (May 2016), p. 9. doi: 10.1186/s40537-016-0043-6. url: https://doi.org/10.1186/s40537-016-0043-6 (cit. on p. 78).

[Wes15] Weston, Jason; Chopra, Sumit and Bordes, Antoine: "Memory Networks". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio, Yoshua and LeCun, Yann. 2015. URL: http://arxiv.org/abs/1410.3916 (cit. on p. 99).

[Wil20] Wilson, Garrett and Cook, Diane J.: "A Survey of Unsupervised Deep Domain Adaptation". In: *ACM Transactions on Intelligent Systems and Technology* 11.5 (July 2020), pp. 1–46. DOI: 10.1145/3400066. URL: http://dx.doi.org/10.1145/3400066 (cit. on p. 48).

[Woj17] Wojke, Nicolai; Bewley, Alex and Paulus, Dietrich: "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2017. DOI: 10.1109/icip.2017.8296962. URL: http://dx.doi.org/10.1109/ICIP.2017.8296962 (cit. on p. 127).

[Wu21] Wu, Zonghan; Pan, Shirui; Chen, Fengwen; Long, Guodong; Zhang, Chengqi and Yu, Philip: "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Trans. Neural Networks Learn. Syst.* 32.1 (2021), pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386. arXiv: 1901.00596 (cit. on pp. 17–20).

[Xia18] Xiao, Bin; Wu, Haiping and Wei, Yichen: "Simple Baselines for Human Pose Estimation and Tracking". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 472–487. DOI: 10.1007/978-3-030-01231-1_29. URL: http://dx.doi.org/10.1007/978-3-030-01231-1_29 (cit. on pp. 122, 138, 234).

[Xu18] Xu, Mingliang; Li, Chunxu; Lv, Pei; Lin, Nie; Hou, Rui and Zhou, Bing: "An Efficient Method of Crowd Aggregation Computation in Public Areas". In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.10 (Oct. 2018), pp. 2814–2825. DOI: 10.1109/tcsvt.2017.2731866. URL: http://dx.doi.org/10.1109/TCSVT.2017.2731866 (cit. on pp. 46, 47).

[Xu22a] Xu, X.; Kostka, G. and Cao, X.: "Information control and public support for social credit systems in china". In: *The Journal of Politics* 84 (4 2022), pp. 2230–2245. DOI: 10.1086/718358 (cit. on p. 5).

[Xu22b] Xu, Yufei; Zhang, Jing; ZHANG, Qiming and Tao, Dacheng: "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K. and Oh, A. Vol. 35. Curran Associates, Inc., 2022, pp. 38571–38584. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/fbb10d319d44f8c3b4720873e4177c65-Paper-Conference.pdf (cit. on pp. 27, 34, 138, 235).

[Xu24] Xu, Yufei; Zhang, Jing; Zhang, Qiming and Tao, Dacheng: "ViTPose++: Vision Transformer for Generic Body Pose Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.2 (Feb. 2024), pp. 1212–1230. DOI: 10.1109/tpami.2023.3330016. URL: http://dx.doi.org/10.1109/TPAMI.2023.3330016 (cit. on p. 27).

[Yan13] Yang, Yi and Ramanan, Deva: "Articulated Human Detection with Flexible Mixtures of Parts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2878–2890. DOI: 10.1109/TPAMI.2012.261 (cit. on p. 112).

[Yan18] Yan, Sijie; Xiong, Yuanjun and Lin, Dahua: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. 2018. DOI: 10.48550/ARXIV.1801.07455. URL: https://arxiv.org/abs/1801.07455 (cit. on pp. 56, 97).

[Yu18] Yu, Bing; Yin, Haoteng and Zhu, Zhanxing: "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. IJCAI-2018. International Joint Conferences on Artificial Intelligence Organization, July 2018. DOI: 10.24963/ijcai.2018/505. URL: http://dx.doi.org/10.24963/ijcai.2018/505 (cit. on p. 97).

[Yu24] Yu, Shoubin; Zhao, Zhongyin; Fang, Haoshu; Deng, Andong; Su, Haisheng; Wang, Dongliang; Gan, Weihao; Lu, Cewu and Wu, Wei: "Regularity Learning via Explicit Distribution Modeling for Skeletal Video Anomaly Detection". In: *IEEE Transactions on Circuits and Systems for Video Technology* (July 2024), pp. 1–1. DOI: 10.1109/tcsvt.2023.3296118. URL: http://dx.doi.org/10.1109/TCSVT.2023.3296118 (cit. on pp. 27, 57, 158, 159, 163, 167).

[Zan20] Zanin, M.; Cheng, X.; Liang, T.; Ling, S.; Zhao, F.; Huang, Z.; Lin, F.; Xia, L.; Jiang, Z. and Wong, S.: "The public health response to the covid-19 outbreak in mainland china: a narrative review". In: *Journal of Thoracic Disease* 12 (8 2020), pp. 4434–4449. DOI: 10.21037/jtd-20-2363 (cit. on p. 5).

[Zan23] Zanfir, Andrei; Zanfir, Mihai; Gorban, Alex; Ji, Jingwei; Zhou, Yin; Anguelov, Dragomir and Sminchisescu, Cristian: "HUM3DIL: Semi-supervised Multi-modal 3D HumanPose Estimation for Autonomous Driving". In: *Proceedings of The 6th Conference on Robot Learning*. Ed. by Liu, Karen; Kulic, Dana and Ichnowski, Jeff. Vol. 205. Proceedings of Machine Learning Research. PMLR, Oct. 2023, pp. 1114–1124. URL: https://proceedings.mlr.press/v205/zanfir23a.html (cit. on p. 37).

[Zen23] Zeng, Xianlin; Jiang, Yalong; Ding, Wenrui; Li, Hongguang; Hao, Yafeng and Qiu, Zifeng: "A Hierarchical Spatio-Temporal Graph Convolutional Neural Network for Anomaly Detection in Videos". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.1 (Jan. 2023), pp. 200–212. DOI: 10.1109/tcsvt.2021.3134410. URL: http://dx.doi.org/10.1109/TCSVT.2021.3134410 (cit. on p. 56).

[Zha15] Zhang, Cong; Li, Hongsheng; Wang, Xiaogang and Yang, Xiaokang: "Cross-scene crowd counting via deep convolutional neural networks". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. DOI: 10.1109/

cvpr.2015.7298684. URL: http://dx.doi.org/10.1109/CVPR.2015.7298684 (cit. on p. 81).

[Zha18a] Zhang, Xinfeng; Yang, Su; Zhang, Xinjian; Zhang, Weishan and Zhang, Jiulong: "Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning". In: *arXiv preprint arXiv:1805.10620* (2018) (cit. on p. 115).

[Zha18b] Zhao, Mingmin; Li, Tianhong; Alsheikh, Mohammad Abu; Tian, Yonglong; Zhao, Hang; Torralba, Antonio and Katabi, Dina: "Through-Wall Human Pose Estimation Using Radio Signals". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. DOI: 10.1109/cvpr.2018.00768. URL: http://dx.doi.org/10.1109/CVPR.2018.00768 (cit. on pp. 38, 39).

[Zha18c] Zhao, Sicheng; Wu, Bichen; Gonzalez, Joseph; Seshia, Sanjit A. and Keutzer, Kurt: "Unsupervised Domain Adaptation: from Simulation Engine to the RealWorld". In: *CoRR* abs/1803.09180 (2018). arXiv: 1803.09180. URL: http://arxiv.org/abs/1803.09180 (cit. on p. 48).

[Zha19] Zhang, Si; Tong, Hanghang; Xu, Jiejun and Maciejewski, Ross: "Graph convolutional networks: a comprehensive review". In: *Comput. Soc. Networks* 6 (Sept. 2019), pp. 1–23. DOI: https://doi.org/10.1186/s40649-019-0069-y (cit. on p. 19).

[Zha20] Zhang, Jinglei and Hou, Yawei: "Image-to-image Translation Based on Improved Cycle-consistent Generative Adversarial Network". In: *Journal of Electronics and Information Technology* 42 (June 2020), pp. 1216–1222. DOI: 10.11999/JEIT190407 (cit. on p. 82).

[Zha21] Zhang, T.; Wang, Q.; Wang, Y.; Bai, G.; Dai, R. and Luo, L.: "Early surveillance and public health emergency responses between novel coronavirus disease 2019 and avian influenza in china: a case-comparison study". In: *Frontiers in Public Health* 9 (2021). DOI: 10.3389/fpubh.2021.629295 (cit. on p. 5).

[Zha22] Zhang, Shaoxiong; Wang, Yunhong; Chai, Tianrui; Li, Annan and Jain, Anil K.: RealGait: Gait Recognition for Person Re-Identification. Jan. 2022. doi: 10.48550/ARXIV.2201.04806. url: https://arxiv.org/abs/2201.04806 (cit. on p. 168).

[Zhe21] Zheng, Sixiao et al.: Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. 2021. arXiv: 2012.15840 [cs.CV] (cit. on p. 27).

[Zhe23] Zheng, Ce; Wu, Wenhan; Chen, Chen; Yang, Taojiannan; Zhu, Sijie; Shen, Ju; Kehtarnavaz, Nasser and Shah, Mubarak: "Deep Learning-based Human Pose Estimation: A Survey". In: *ACM Computing Surveys* 56.1 (Aug. 2023), pp. 1–37. doi: 10.1145/3603618. url: http://dx.doi.org/10.1145/3603618 (cit. on p. 34).

[Zho14] Zhou, Bolei; Tang, Xiaoou; Zhang, Hepeng and Wang, Xiaogang: "Measuring Crowd Collectiveness". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.8 (Aug. 2014), pp. 1586–1599. doi: 10.1109/tpami.2014.2300484. url: http://dx.doi.org/10.1109/TPAMI.2014.2300484 (cit. on pp. 46, 47).

[Zho19] Zhou, Kaiyang; Yang, Yongxin; Cavallaro, Andrea and Xiang, Tao: "Omni-Scale Feature Learning for Person Re-Identification". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019. doi: 10.1109/iccv.2019.00380. url: http://dx.doi.org/10.1109/ICCV.2019.00380 (cit. on p. 133).

[Zho23] Zhou, Xiaolong; Jin, Tian; Dai, Yongpeng; Song, Yongkun and Qiu, Zhifeng: "MD-Pose: Human Pose Estimation for Single-Channel UWB Radar". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 5.4 (Oct. 2023), pp. 449–463. doi: 10.1109/tbiom.2023.3265206. url: http://dx.doi.org/10.1109/TBIOM.2023.3265206 (cit. on p. 38).

[Zhu17] Zhu, Jun-Yan; Park, Taesung; Isola, Phillip and Efros, Alexei A.: "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *2017 IEEE International*

Conference on Computer Vision (ICCV). 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244 (cit. on pp. 24, 48, 49, 82, 91, 93, 94, 233).

[Zhu22] Zhu, Yean; Lu, Wei; Zhang, Ruoqi; Wang, Rui and Robbins, Dan: "Dual-channel cascade pose estimation network trained on infrared thermal image and groundtruth annotation for real-time gait measurement". In: *Medical Image Analysis* 79 (July 2022), p. 102435. DOI: 10.1016/j.media.2022.102435. URL: http://dx.doi.org/10.1016/j.media.2022.102435 (cit. on p. 35).

Own Publications

This section serves as a comprehensive index of the author's publications, which can be organized into four distinct topics. These topics encompass various areas, including:

- Human Pose Estimation and Tracking – Explored in [2], [9].
- Anomaly Detection and Behavior Analysis – Covered in [1], [3], [5], [12], [13], [14].
- Domain Adaptation – Addressed in [6].
- Crowd Monitoring – Investigated in [4], [8], [10], [11].

The publications falling into the field of Crowd Monitoring, along with the publication [7], have broader relevance to this dissertation.

[1] Golda, Thomas: "Image-based Anomaly Detection within Crowds". In: *Proceedings of the 2018 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer, M. Taphanel.* Vol. 40. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2019, pp. 11–24. DOI: 10.5445/IR/1000097082.

[2] Golda, Thomas; Kalb, Tobias; Schumann, Arne and Beyerer, Jürgen: "Human Pose Estimation for Real-World Crowded Scenarios". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).* 2019, pp. 1–8. DOI: 10.1109/AVSS.2019.8909823.

[3] Golda, Thomas; Murzyn, Nils; Qu, Chengchao and Kroschel, Kristian: "What goes around comes around: Cycle-Consistency-based Short-Term Motion Prediction for Anomaly Detection using Generative Adversarial Networks". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019, pp. 1–8. doi: 10.1109/AVSS.2019.8909853.

[4] Du, Dawei et al.: "VisDrone-CC2020: The Vision Meets Drone Crowd Counting Challenge Results". In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Bartoli, Adrien and Fusiello, Andrea. Cham: Springer International Publishing, 2020, pp. 675–691. doi: 10.1007/978-3-030-66823-5_41.

[5] Golda, Thomas: "Part Affinity Field based Activity Recognition". In: *Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory / by Jürgen Beyerer, Tim Zander (eds.)* Vol. 45. Karlsruher Schriften zur Anthropomatik. KIT Scientific Publishing, 2020, pp. 53–67. doi: 10.5445/IR/1000126337.

[6] Golda, Thomas; Blattmann, Andreas; Metzler, Jürgen and Beyerer, Jürgen: "Image domain adaption of simulated data for human pose estimation". In: *Artificial Intelligence and Machine Learning in Defense Applications II*. Ed. by Dijk, Judith. Vol. 11543. International Society for Optics and Photonics. SPIE, 2020, pp. 112–127. doi: 10.1117/12.2573888. url: https://doi.org/10.1117/12.2573888.

[7] Cormier, Mickael; Röpke, Fabian; Golda, Thomas and Beyerer, Jürgen: "Interactive Labeling for Human Pose Estimation in Surveillance Videos". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021, pp. 1649–1658. doi: 10.1109/ICCVW54120.2021.00190.

[8] Fraunhofer IOSB; Golda, Thomas; Krempel, Erik; Metzler, Jürgen and Westerkampf, Kai: Verbundprojekt: Sicherheit in städtischen Umgebungen: Crowd-Monitoring, Prädiktion und Entscheidungsunterstützung (S2UCRE). German. Tech. rep. Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB), 2021.

[9] Golda, Thomas: ""Let's get ready to bundle!: Crowd-level Human Keypoint Tracking". In: *Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer; T. Zander.* Vol. 51. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2021, pp. 67–81. doi: 10.5445/IR/1000135196.

[10] Golda, Thomas; Krüger, Florian and Beyerer, Jürgen: "Temporal Extension for Encoder-Decoder-based Crowd Counting Approaches". In: *2021 17th International Conference on Machine Vision and Applications (MVA).* 2021, pp. 1–5. doi: 10.23919/MVA51890.2021.9511351.

[11] Kramer, Jan Calvin; Golda, Thomas; Hansert, Jonas and Schlegel, Thomas: "Improving Temporal Consistency in Aerial Based Crowd Monitoring Using Bayes Filters". In: *UR-AI 2021, The Upper-Rhine Artificial Intelligence Symposium : Artificial Intelligence - Application in life sciences and beyond, Kaiserslautern, 27th October 2021. Ed.: K-H. Schäfer.* Upper-Rhine Artificial Intelligence Symposium. 2021 (Kaiserslautern, Deutschland, Oct. 27, 2021–). Hochschule Kaiserslautern, 2021, pp. 224–233.

[12] Golda, Thomas; Guaia, Deborah and Wagner-Hartl, Verena: "Perception of Risks and Usefulness of Smart Video Surveillance Systems". In: *Applied Sciences* 12.20 (2022). doi: 10.3390/app122010435. url: https://www.mdpi.com/2076-3417/12/20/10435.

[13] Golda, Thomas; Thiemich, Johanna; Cormier, Mickael and Beyerer, Jürgen: "For the Sake of Privacy: Skeleton-Based Salient Behavior Recognition". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 3983–3987. doi: 10.1109/ICIP46576.2022.9897358.

[14] Golda, Thomas; Cormier, Mickael and Beyerer, Jürgen: "Intelligente Bild- und Videoauswertung für die Sicherheit". In: *Handbuch Polizeimanagement: Polizeipolitik – Polizeiwissenschaft – Polizeipraxis*. Ed. by Wehe, Dieter and Siller, Helmut. Wiesbaden: Springer Fachmedien Wiesbaden, 2023, pp. 1487–1507. doi: 10.1007/978-3-658-34388-0_87. url: https://doi.org/10.1007/978-3-658-34388-0_87.

Supervised Student Theses

This section provides an overview of the supervised bachelor's and master's theses conducted in the course of this doctorate. These theses can be categorized into different subtopics that have contributed to this work. In the context of this research, the theses can be divided into two main areas of focus. The first area pertains to aspects related to human pose estimation and has been explored within the following theses: [1], [2], [4], [7], [11]. On the other hand, the second area of focus involves work related to behavior analysis, which has been examined in the following theses: [5], [6], [13], [14], [16]. The remaining theses, [3], [8], [9], [10], [12], [15], have a broader relevance to this work as they predominantly focus on crowd counting and crowd density estimation, with the exception of [9], which deals with data labeling — a topic not covered in this thesis.

[1] Dissert, Thomas: "Crowd level Person Pose Estimation". Bachelor's Thesis. Karlsruhe Institute of Technology, Oct. 2018 (cit. on p. 213).

[2] Blattmann, Andreas: "Multi Person Pose Estimation using Synthetically Generated Data". Master's Thesis. Karlsruhe Institute of Technology, July 2019 (cit. on p. 213).

[3] Bleymehl, Tobias: "Dichtebasiertes Zählen von Menschen in Bilddaten". Bachelor's Thesis. Karlsruhe Institute of Technology, Feb. 2019 (cit. on p. 213).

[4] Kalb, Tobias: "Optimization of Human Body Pose Estimation for Crowd Applications". Master's Thesis. Karlsruhe Institute of Technology, May 2019 (cit. on p. 213).

[5] Katovich, Kristina: "Artifical Data for Activity Recognition". Bachelor's Thesis. Karlsruhe Institute of Technology, Nov. 2019 (cit. on p. 213).

[6] Murzyn, Nils: "GAN based Anomaly Detection". Master's Thesis. Karlsruhe Institute of Technology, May 2019 (cit. on p. 213).

[7] Becker, Oliver: "Multi-Keypoint Tracking auf synthetischen Menschenmengen". Bachelor's Thesis. Karlsruhe Institute of Technology, Oct. 2020 (cit. on p. 213).

[8] Krüger, Florian: "Crowd Density Estimation from Aerial Imagery". Master's Thesis. Karlsruhe Institute of Technology, Dec. 2020 (cit. on p. 213).

[9] Arndt, Fabian: "Entwicklung einer webbasierten Preprocessing Anwendung im Bereich des Machine Learnings". Bachelor's Thesis. Hochschule Furtwangen, Mar. 2021 (cit. on p. 213).

[10] Kienle, Claudius: "Portierung videobasierter Menschenmengenanalyse auf ein Edge Device". Bachelor's Thesis. Karlsruhe Institute of Technology, Jan. 2021 (cit. on p. 213).

[11] Kong, Xiaoyan: "Generative Adversarial Network based Image Domain Adaption for Multi Person Pose Estimation". Master's Thesis. Karlsruhe Institute of Technology, June 2021 (cit. on p. 213).

[12] Kramer, Calvin: "Crowd Monitoring in Aerial Imagery". Bachelor's Thesis. Hochschule Karlsruhe, Mar. 2021 (cit. on p. 213).

[13] Thiemich, Johanna: "Skeleton Based Detection of Salient Human Behavior in Urban Scenarios". Master's Thesis. Karlsruhe Institute of Technology, Nov. 2021 (cit. on p. 213).

[14] Guaia, Deborah: "Einflussfaktoren auf die Akzeptanz bildbasierter Anomalieerkennung in Menschenmengen". Bachelor's Thesis. Hochschule Furtwangen, Apr. 2022 (cit. on p. 213).

[15] Sänger, Jann: "Evaluation of Density Based Crowd Counting Approaches". Bachelor's Thesis. Hochschule Karlsruhe, Mar. 2023 (cit. on p. 213).

[16] Hohloch, David: “Skeleton-based Behavior Analysis”. Bachelor’s Thesis. Karlsruhe Institute of Technology, Jan. 2024 (cit. on p. 213).

List of Figures

List of Tables

Acronyms

AE	Autoencoder
AI Act	Artificial Intelligence Act
AP	average precision
AR	average recall
AUC	Area under the curve
AUC-PR	Area under the Precision-Recall curve
AUC-ROC	Area under the Receiver Operator Characteristic curve
BinAE	Binary Autoencoder
CA	Channel Attention
CaWa	Cannstatter Wasen
CCTV	Closed-Circuit Television
cGAN	conditional Generative Adversarial Network
CHAD	Charlotte Anomaly Dataset

CI	Crowd Index
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
COCO	Common Objects in Context
COVID-19	Coronavirus disease 2019
CR	Crowd Ratio
CrowDPB	Crowded Dataset for Human Pose Estimation and Behavior Analysis
CYCADA	Cycle-Consistent Adversarial Domain Adaptation
DA	Domain Adaption
DMAD	Diversity-Measurable Anomaly Detection
DPMM	Dirichlet Process Mixture Model
EER	Equal Error Rate
EU	European Union
FID	Fréchet inception distance
FN	False Negative
FNR	False Negative Rate
FP	False Positive

FPR	False Positive Rate
GAN	Generative Adversarial Network
GCI	Graph Crowd Index
GCNN	Graph Convolutional Neural Network
GDPR	General Data Protection Regulation
GEPC	Graph Embedded Pose Clustering
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
GTA	Grand Theft Auto
HeR	Hessischer Rundfunk
HGH	Hafengeburtstag Hamburg
HPE	Human Pose Estimation
HPERL	3D Human Pose Estimation from RGB and LiDAR
HRNet	High-Resolution Net
HSTGCNN	Hierarchical Spatio-Temporal Graph Convolutional Neural Network
HUM3DIL	Semi-supervised Multi-modal 3D Human Pose Estimation for Autonomous Driving

HuPR	Human Pose with Millimeter Wave Radar
IoU	Intersection over Union
IPHPDT	Identity-Preserved Human Posture Detection in Thermal Images
JTA	Joint Track Auto
LiDAR	Light Imaging, Detection and Ranging
LSTM	Long Short-Term Memory
mAP	mean average precision
mAR	mean average recall
ME	Motion Embedder
MemAE	Memory-augmented Autoencoder
MixAMoR	Mixamo Anomalous Movement Recognition
MLP	Multilayer Perceptron
MoPRL	Motion Prior Regularity Learner
MOT	Multi-Object Tracking
MPED-RNN	Message-Passing Encoder-Decoder Recurrent Neural Network
MPII	MPII Human Pose Dataset
MPPE	Multi-Person Pose Estimation

MSE	Mean Squared Error
nGCI	Normalized Graph Crowd Index
NIR	near infrared
NLP	Natural Language Processing
NMS	Non-maximum Suppression
NN	Neural Network
OF	optical flow
OKS	Object Keypoint Similarity
OSNet	Omni-Scale Network
PCK	Probability of Correct Keypoint
PCKh	Probability of Correct Keypoint (head)
PR	Precision-Recall Curve
PSAP	Public Safety Answering Point
ReID	re-identification
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic

SAA-STGCAE	Spatial Temporal Self-attention Augmented Graph Convolutional Autoencoder
SAHI	Slicing Aided Hyper Inference
SARS	Severe acute respiratory syndrome
SBA	Skeleton-based Behavior Analysis
SBAD	Skeleton-based Anomaly Detection
SBAR	Skeleton-based Action Recognition
SCAPE	Shape Completion and Animation of People
SCS	Social Credit System
sGCI	Standard Graph Crowd Index
SHT	ShanghaiTech
SHTC	ShanghaiTech Campus
SHTC-HR	Human-Related ShanghaiTech Campus
SMPL	Skinned Multi-Person Linear Model
SPPE	Single-Person Pose Estimation
STemGAN	Spatio-Temporal Generative Adversarial Network
ST-GCAE	Spatio-temporal Graph Autoencoder
ST-GCN	Spatial-Temporal Graph Convolutional Network

STG-NF	Spatio-Temporal Graph Normalizing Flows
STT	Spatial-Temporal Transformer
Swin	Swin Transformer
SyMPose	Synthetic Semantic Maps and Human Poses
TIR	thermal infrared
TL	transfer learning
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TPU	Tensor Processing Unit
t-SNE	t-Distributed Stochastic Neighbor Embedding
VAD	Video-based Anomaly Detection
VFP290K	Vision-based Fallen Person
ViT	Vision Transformer
WE	World Expo
YOLO	You Only Look Once

Glossary

1080p set of high-definition video modes characterized by 1,920 pixels displayed across the screen horizontally and 1,080 pixels down the screen vertically; also known as *Full HD* or *FHD*

480p set of video modes characterized by 480 pixels displayed down the screen vertically

720p set of high-definition video modes characterized by 1,280 pixels displayed across the screen horizontally and 720 pixels down the screen vertically; also known as *HD*

HRNet Neural Network (NN) architecture for HPE that is able to maintain high-resolution representations through the whole process. It starts from a high-resolution subnetwork as the first stage, and gradually adds high-to-low resolution subnetworks one by one to form more stages, and finally connects the multi-resolution subnetworks in parallel [Sun19]

AlexNet first CNN that used GPU to boost performance [Kri12]

AlphaPose top-down human pose estimator by Fang et al. [Fan23]

BERT Bidirectional Encoder Representations from Transformers [Dev18]

CaWa18 internal dataset collected during the CaWa 2018

ConvNeXt modern CNN architecture

Crowd Aggregation motion based measure to characterize crowds

Crowd Collectiveness measure to characterize a given crowd based on its intrinsic motion and movements

CrowDPB Crowded Dataset for Human Pose Estimation and Behavior Analysis is a collection of video sequences and frames created for this thesis with the purpose of evaluating performance of Human Pose Estimation, Domain Adaption, Video-based Anomaly Detection, and Skeleton-based Anomaly Detection

CrowdPE internal multi-purpose dataset consisting of various single frames and sequences showing crowded situations on different events

CrowdPose diverse dataset for human pose estimation in crowded scenes, crucial for training algorithms in complex environments.

Cycle-GAN neural network architecture consisting out of two GAN frameworks used for various task where different images have to be translated into each other [Zhu17]

DualHeadAE Two-headed AE consisting of one encoder and two decoders, one for normal and the other for abnormal inputs.

FBX proprietary file format providing interoperability between different digital content creation applications

Fraunhofer IOSB Fraunhofer-Institute of Optronics, System Technologies and Image Exploitation IOSB

HeR19 internal dataset containing videos provided in 2019 by HeR

HGH18 internal dataset collected during the Hafengeburtstag Hamburg 2018

HGH18-DA subset of HGH18 without annotations but representing the target domain in DA

HGH18-HPE subset of HGH18 annotated with keypoints and bounding boxes of humans

ImageNet image database consisting of millions of images, which is the basis for the training of most large-scale models

JFT-300M internal dataset from Google for training image classification models. It contains 300 million images with approximately 375 million labels, which are selected by an algorithm to maximize precision for the chosen images [Sun17]

LSGAN similar to the VanillaGAN, but using the least squares loss

macroscopic refers to methods in the scope of anomaly detection that are *human-centered*

microscopic refers to methods in the scope of anomaly detection that are *holistic*

MurzGAN GAN-based approach for VAD [Gol19c]

OpenPose bottom-up human pose estimator by Cao et al. [Cao21]

pix2pix cGAN architecture that was developed by Isola et al. [Iso17]. Unlike VanillaGAN which uses only real data and noise to learn and generate images, cGAN uses real data, noise as well as labels to generate images.

PoseTrack dataset for HPE mostly focusing on videos

SimpleBaselines collection of human pose estimation networks provided by Xiao et al. [Xia18]

StrongSORT deep learning based framework for tracking objects introduced by Du et al. [Du23]

Transformer	NN architecture initially presented for the task of NLP that is applied to many other fields and is the basis of many foundation models
U-Net	U-Net is a widely known encoder-decoder-network that was firstly applied for the segmentation of neuronal structures [Ron15]
VanillaGAN	Reference GAN architecture as initially proposed by Goodfellow et al. [Goo14]
VGG16	neural network architecture typically used for image classification, and today still used as feature extractor for various applications
ViTPose	human pose estimation architecture build upon the Vision Transformer [Xu22b]
WiFi	is a family of wireless network protocols based on the IEEE 802.11 family of standards, which are commonly used for local area networking of devices and Internet access, allowing nearby digital devices to exchange data by radio waves
WorldExpo '10	public dataset collected during the World Expo 2010 in Shanghai, China
YOLO	object detection framework that consists of multiple versions

A Appendix

A.1 SyMPose

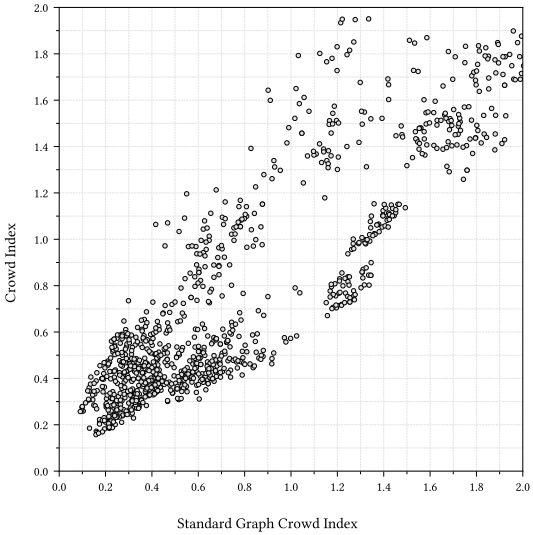


Figure A.1: Visual comparison of CI and sGCI for different images. This illustration is based on a uniform random draw from SyMPose. Each point corresponds to a frame taken from SyMPose. The figure shows, that with increasing CI the sGCI spreads.

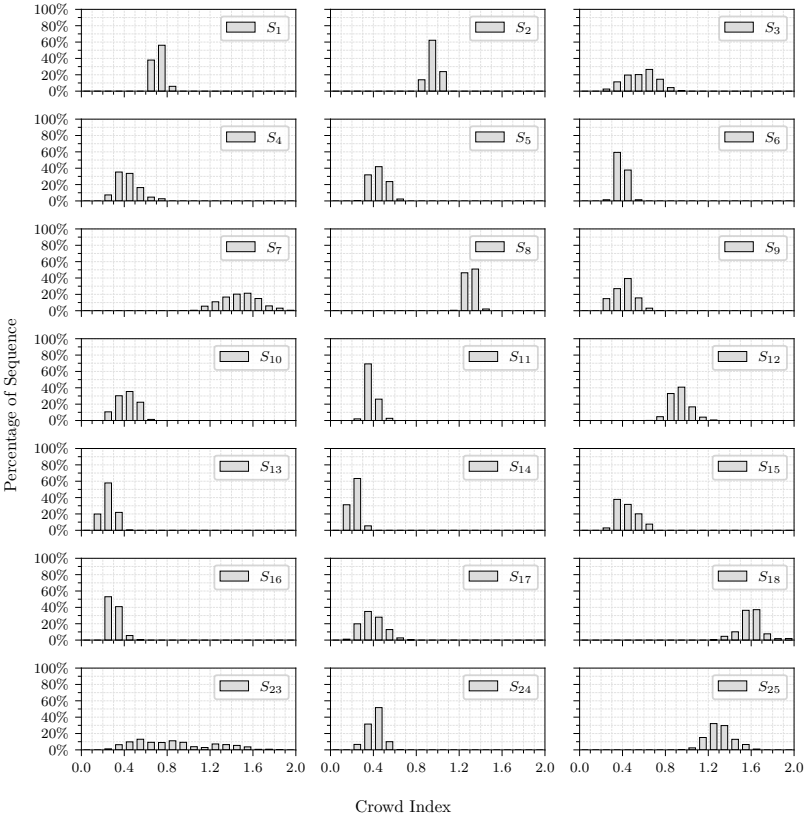


Figure A.2: Overview over all index CI distributions for each sequence contained in the final SyMPose dataset. The Crowd Index stays for most of the time near to a value of **1.0** with only few stronger outliers as there are e.g., in S_8 and S_{18}.

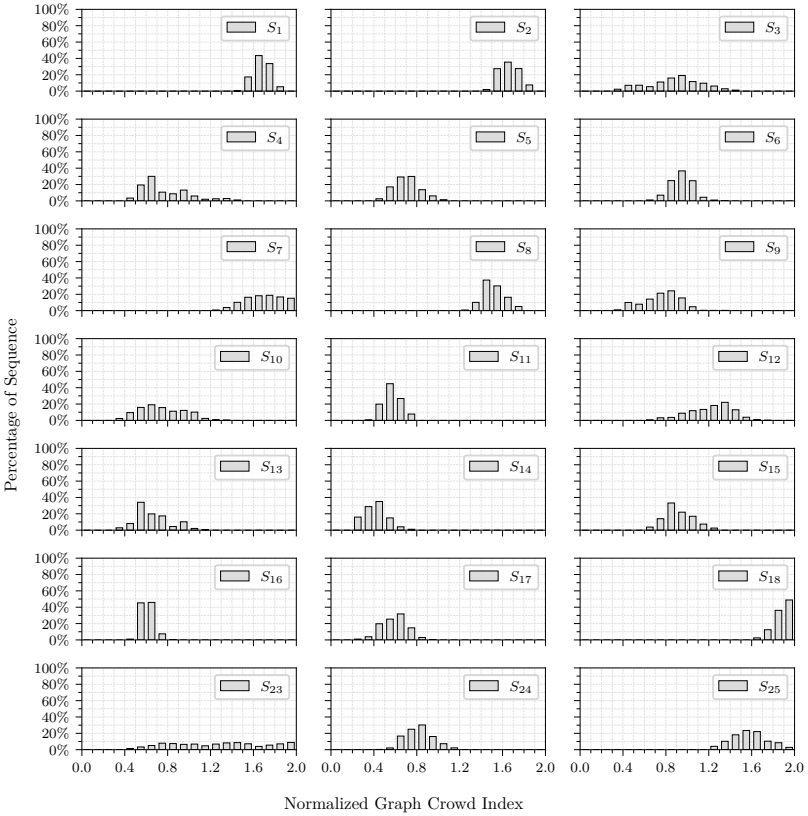


Figure A.3: Overview over all index nGCI distributions for each sequence contained in the final SyMPose dataset. It is obvious that the nGCI results in a shift towards higher values compared to those from CI.

Figure A.4: Exemplary frames from SyMPose. These are corresponding source domain images for the results in Figure 5.8 and Figure 5.10.

A.2 CrowDPB: Human Examples

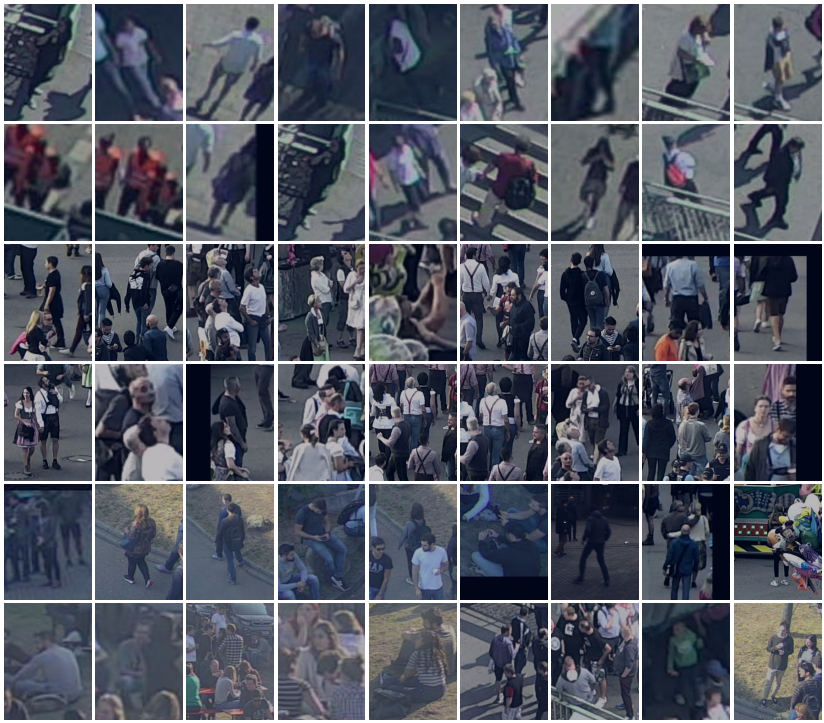


Figure A.5: Random examples from CrowDPB. The upper two rows show samples from HGH18, the middle two rows from CaWa18, and the lower two rows from CrowdPE.

A.3 Crowd Measures: Examples and Supplements

A.3.1 Example: Crowd Index

For a better understanding of the Crowd Index a simplified example is given in Figure A.6a. The reference person 0 has two interfering individuals shown in green (person 1) and blue (person 2), which partially colide with the bounding box of person 0. Based on this scenario, the resulting Crowd Ratio is

$$\xi_0 = \frac{n_0^1 + n_0^2}{n_0^0} = \frac{5+5}{15} = \frac{2}{3} \approx 0.6667 \tag{A.1}$$

This example already shows a major flaw of the CI: Person 1 and 2 are border cases, whereas person 0 is an inner case. Albeit the relatively high Crowd Ratio for person 0, the overall effect on the CI is rather low, since for such scenario the border cases exceed the inner cases.

$$\xi_1 = \frac{n_1^0 + n_1^2}{n_1^1} = \frac{5+0}{15} = \frac{1}{3} \approx 0.3333 \tag{A.2}$$

$$\xi_2 = \frac{n_2^0 + n_2^1}{n_2^2} = \frac{2+0}{15} = \frac{2}{15} \approx 0.1333 \tag{A.3}$$

As by definition, the resulting CI is

$$f_{\mathrm{CI}} = \frac{1}{3}\sum_{i=0}^{2} \xi_i = \frac{1}{3}\left(\frac{2}{3} + \frac{1}{3} + \frac{2}{15}\right) = \frac{17}{45} \approx 0.3778 \tag{A.4}$$

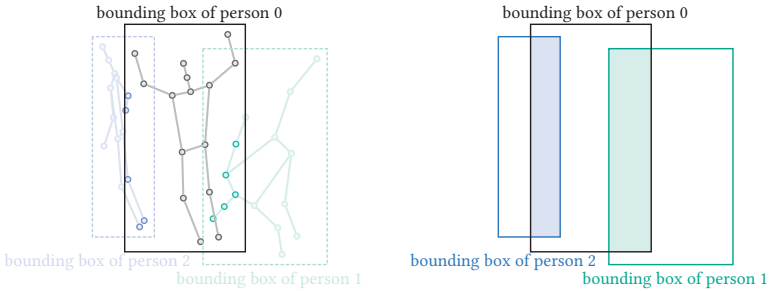
In contrast, the corresponding MOT measure yields $f_{\mathrm{MOT}} = 3$, which is less informative than the CI since the value does not give any insight in the overall

relative spatial distribution of the individuals.

$$\begin{aligned} A_0 &= w \cdot h = 12 \cdot 23 = 276 \\ A_1 &= w \cdot h = 13 \cdot 22 = 286 \\ A_2 &= w \cdot h = 7 \cdot 20 = 140 \end{aligned} \tag{A.5}$$

$$\begin{aligned} \psi_{01} &= \frac{276}{286} \approx 0.9650 \\ \psi_{02} &= \frac{140}{276} \approx 0.5072 \end{aligned} \tag{A.6}$$

$$\begin{aligned} J_{01} &= \frac{252}{276 + 286 - 252} = \frac{252}{310} \approx 0.8129 \\ J_{02} &= \frac{114}{276 + 140 - 114} = \frac{114}{302} \approx 0.3775 \end{aligned} \tag{A.7}$$



(a) Example with Keypoints **(b)** Example with Bounding Boxes

Figure A.6: Exemplary illustration of the original and generalized Crowd Ratio. On the left part of this figure, the circles indiciate keypoints corresponding to three different persons. Keypoints within the black bounding box of the reference person are drawn in a dark color, whereas keypoints outside the bounding box and hence without effect on the Crowd Ratio are shown in a lighter tone. In addition, the bounding boxes for person 1 and 2 are indicated as well by the colored dashed rectangles. The right hand side shows the identical scenario as given on the left but with no keypoints available. This can be seen as a more general case compared to the original CI.

The resulting area is hence $w \cdot h = 3.5 \cdot 4.5 = 15.75$ for each boundingbox and therefore $\psi_{ik} = 1.0$ for every $i, k \in \{0,1,2\}$.

A.3.2 Graph Crowd Index

This section provides some additional information and mathematical considerations that explain or underline the properties of the GCI.

Proposition A.1. *$J_{ik} \leq \psi_{ik}$ for $i,k \in \mathcal{I}$*

Proof. w.l.o.g. let $|A_i| \leq |A_k|$ with $|A_i|, |A_k| > 0$. Furthermore, let $|A_i| \coloneqq a_1 a_2$ and $|A_k| \coloneqq b_1 b_2$.
There are three cases to consider:

(C0): $A_i \cap A_k = A_i$

(C1): $A_i \cap A_k = \varnothing$

(C2): $A_i \cap A_k = X \subset A_i$ with $X \neq \varnothing$

Case (C1):
$J_{ik} = \frac{|A_i \cap A_k|}{|A_i \cup A_k|} = \frac{|A_i|}{|A_k|} = \frac{a_1 a_2}{b_1 b_2} = \psi_{ik}$

Case (C2):
$J_{ik} = \frac{|A_i \cap A_k|}{|A_i \cup A_k|} = \frac{|\varnothing|}{|A_i| + |A_k|} = \frac{0}{a_1 a_2 + b_1 b_2} < \frac{a_1 a_2}{b_1 b_2} = \psi_{ik}$

Case (C3):
From the given pre-conditions results that $|X| < |A_i| \leq |A_k|$, and therefore
$J_{ik} = \frac{|A_i \cap A_k|}{|A_i \cup A_k|} = \frac{|X|}{|A_i| + |A_k| - |X|} = \frac{x_1 x_2}{a_1 a_2 + b_1 b_2 - x_1 x_2} < \frac{x_1 x_2}{b_1 b_2} < \frac{a_1 a_2}{b_1 b_2} = \psi_{ik}$ □

Proposition A.2. $s_{i,nGCI} \geq \frac{s_{i,sGCI}}{\sqrt{n}}$

Proof. w.l.o.g. let $|A_i| \geq 1$, i.e.,..., and let $\mathbf{a} \coloneqq (|A_0|, \dots, |A_{n-1}|)$

$$s_{i,\text{nGCI}} \overset{\text{def}}{=} \sqrt{\frac{|A_i|}{\frac{1}{n}\sum_{j=0}^{n-1}|A_j|}} = \sqrt{n} \cdot \sqrt{\frac{|A_i|}{\sum_{j=0}^{n-1}|A_j|}} = \sqrt{n} \cdot \frac{\sqrt{|A_i|}}{\sqrt{\sum_{j=0}^{n-1}|A_j|}}$$
$$\geq \sqrt{n} \cdot \frac{\sqrt{|A_i|}}{\sum_{j=0}^{n-1}\sqrt{|A_j|}} = \frac{s_{i,\text{sGCI}}}{\sqrt{n}}$$

□

Proposition A.3. $f_{sGCI} \leq \lambda \cdot f_{nGCI}$

Proof. w.l.o.g. let $|A_i| \geq 1$, i.e.,..., $\lambda \coloneqq \frac{L_{\text{sGCI}}}{L_{\text{nGCI}}}$ and let $\mathbf{a} \coloneqq (|A_0|, \dots, |A_{n-1}|)$

$$
\begin{aligned}
f_{\text{sGCI}} &= L_{\text{sGCI}} \cdot \sum_{i=0}^{n-1} c_i \cdot s_{i,\text{sGCI}} \\
&= L_{\text{sGCI}} \cdot \sum_{i=0}^{n-1} c_i \cdot s_{i,\text{sGCI}} \cdot \frac{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}}{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}} \\
&\leq L_{\text{sGCI}} \cdot \sum_{i=0}^{n-1} c_i \cdot \sqrt{n} \cdot s_{i,\text{nGCI}} \cdot \frac{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}}{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}} \\
&= \sqrt{n} \cdot L_{\text{sGCI}} \cdot \sum_{i=0}^{n-1} \frac{c_i \cdot s_{i,\text{nGCI}} \cdot \sum_{j=0}^{n-1} s_{j,\text{nGCI}}}{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}} \\
&= \lambda \cdot \sqrt{n} \cdot L_{\text{nGCI}} \cdot \sum_{i=0}^{n-1} \frac{c_i \cdot s_{i,\text{nGCI}}}{\sum_{j=0}^{n-1} s_{j,\text{nGCI}}} \cdot \sum_{j=0}^{n-1} s_{j,\text{nGCI}} \\
&= \lambda \cdot \sqrt{n} \cdot L_{\text{nGCI}} \cdot \sum_{i=0}^{n-1} \frac{c_i \cdot \sqrt{|A_i|}}{\sum_{j=0}^{n-1} \sqrt{|A_j|}} \cdot \sum_{j=0}^{n-1} \sqrt{|A_j|} \\
&= \lambda \cdot \sqrt{n} \cdot f_{\text{nGCI}} \cdot \sum_{j=0}^{n-1} \sqrt{|A_j|} \\
&\leq \lambda \cdot f_{\text{nGCI}}
\end{aligned}
$$

□

A.4 Behavior Analysis Datasets

Figure A.7: Exemplary frames from UCSD Pedestrian dataset that was used as a basis for the evaluations of MurzGAN in [Gol19c]. The dataset is characterized by a very low resolution and a static setup that consists of just one view. Furthermore, the image material is recorded in grayscale. According to Sharif et al. [Sha22] its two subsets *Ped1* and *Ped2* are still among the most frequently used anomaly detection datasets. As of mid-2024, they lead the list of used video anomaly data sets.

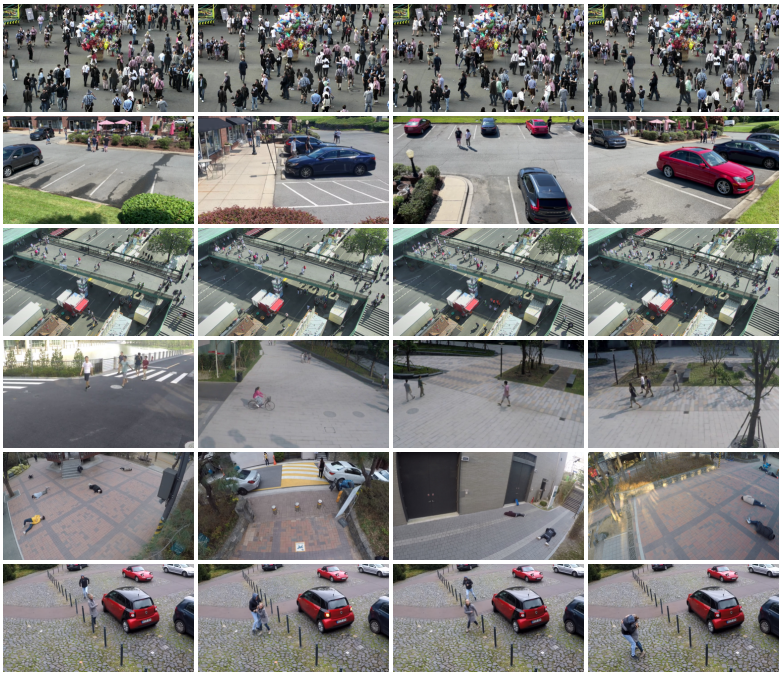


Figure A.8: Exemplary frames from the different datasets. Each row contains four images per dataset, from top to bottom: CaWa18, CHAD, HGH18, SHTC, VFP290K, HeR19.

Table A.1: Results for different blocks or layers of different semantic feature extractors.

Layer / Block	ViT-B16	ViT-B32	VGG16
1	0.6530 ± 0.1021	0.6297 ± 0.0755	0.6508 ± 0.0816
2	0.6311 ± 0.1057	0.6191 ± 0.1047	0.6543 ± 0.0806
3	0.6138 ± 0.1078	0.6114 ± 0.1080	0.6458 ± 0.0778
4	0.5947 ± 0.1061	0.5952 ± 0.1081	0.6700 ± 0.0854
5	0.5810 ± 0.0889	0.5808 ± 0.1082	0.6703 ± 0.0852
6	0.4971 ± 0.0609	0.5644 ± 0.1040	0.6630 ± 0.0882
7	0.5207 ± 0.0557	0.5678 ± 0.1046	0.6850 ± 0.0888
8	0.5368 ± 0.0475	0.5790 ± 0.1014	0.6842 ± 0.0911
9	0.5445 ± 0.0405	0.6033 ± 0.1115	0.6635 ± 0.0844
10	0.5564 ± 0.0388	0.6147 ± 0.1015	0.6852 ± 0.0887
11	0.5764 ± 0.0605	0.6180 ± 0.1087	0.6897 ± 0.0890
12	0.5765 ± 0.0646	0.5962 ± 0.1088	0.6789 ± 0.0854
13	—	—	0.6750 ± 0.1086

Table A.2: Results for different blocks or layers of different semantic feature extractors.

Layer / Block	Swin-T	Swin-B	ConvNeXt-t	ConvNeXt-s
1	0.6049 ± 0.0877	0.6038 ± 0.0825	0.6030 ± 0.0959	0.5795 ± 0.0909
2	0.5842 ± 0.0911	0.5669 ± 0.0828	0.5801 ± 0.0907	0.5625 ± 0.0842
3	0.4773 ± 0.0591	0.4917 ± 0.0304	0.5230 ± 0.0625	0.5271 ± 0.0546
4	0.4929 ± 0.0432	0.4880 ± 0.0354	0.5611 ± 0.0818	0.5383 ± 0.0509

A.5 Domain Adaptation

A.5.1 Hafengeburtstag Hamburg 2018

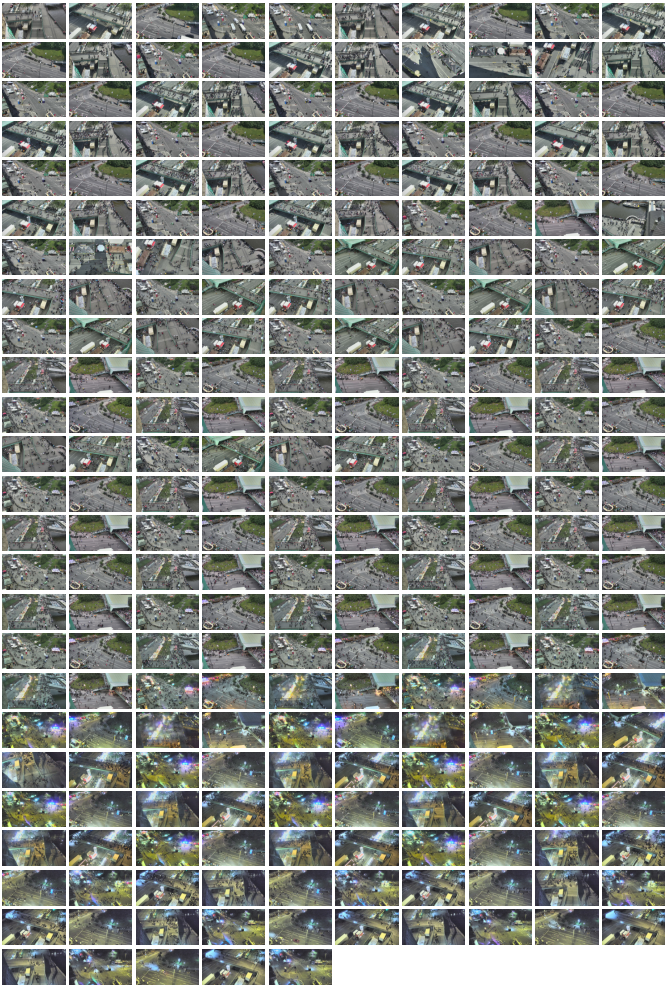


Figure A.9: Overview over sequences contained within the HGH18 dataset sorted by ascending recording time from top-left to bottom-right.

Table A.3: Mean and standard deviation values for different own datasets and the corresponding color channels, where each channel takes values between 0 and 1. These values are the basis for data preprocessing. Apparently, every dataset shows a slightly different characteristic. The HGH18 datasets are slightly lighter compared to the other datasets, having an overall higher amount of green color which comes from the meadow which is visible in most of the frames. CaWa18 on the other hand is slightly darker and shows a similar color distribution as SyMPose-Urban.

Dataset	blue	green	red
SyMPose	0.3953 ± 0.2051	0.4231 ± 0.2379	0.4355 ± 0.2538
SyMPose-Urban	0.3856 ± 0.1776	0.3737 ± 0.1878	0.3709 ± 0.1934
HGH18-DA	0.4299 ± 0.1996	0.4534 ± 0.2016	0.4358 ± 0.1984
HGH18-HPE	0.4674 ± 0.2210	0.4864 ± 0.2371	0.4577 ± 0.2403
CaWa18	0.3900 ± 0.2329	0.3934 ± 0.2359	0.3914 ± 0.2354

A.5.2 Additional Qualitative Results

Figure A.10: Exemplary training patches that were generated for the adaptation to HGH18 target domain. The St. Pauli Pier is characterized by large areas that covered by oxidized copper plates. These are used for the roof, but also for other structures visible in the original video material. The copper elements are complemented by the matching greenish mesh that was attached to the tower on which the cameras were mounted and which is also visible in many samples.