Attribute-Based Person Retrieval in Multi-Camera Networks

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

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

genehmigte

Dissertation

von

M.Sc.

Andreas Heinrich Specker

aus Kirchheim unter Teck

Tag der mündlichen Prüfung:	15.02.2024
Erster Gutachter:	Prof. Dr.-Ing. Jürgen Beyerer
Zweiter Gutachter:	Prof. Dr.-Ing. Rainer Stiefelhagen

Abstract

Attribute-based person retrieval is a crucial component in various real-world applications, including surveillance, retail, and smart cities. Contrary to image-based person identification or re-identification, individuals are searched for based on descriptions of their soft biometric attributes, such as gender, age, and clothing colors. For instance, attribute-based person retrieval enables law enforcement agencies to efficiently search enormous amounts of surveillance footage gathered from multi-camera networks to locate suspects or missing persons.

This thesis presents a novel deep learning framework for attribute-based person retrieval. The primary objective is to research a holistic approach that is suitable for real-world applications. Therefore, all necessary processing steps are covered. Pedestrian attribute recognition serves as the base framework to address attribute-based person retrieval in this thesis. Various design characteristics of pedestrian attribute recognition approaches are systematically examined toward their suitability for attribute-based person retrieval. Following this analysis, novel techniques are proposed and discussed to further improve the performance. The PARNorm module is introduced to normalize the model's output logits across both the batch and attribute dimensions to compensate for imbalanced attributes in the training data and improve person retrieval performance simultaneously. Strategies for video-based pedestrian attribute recognition are explored, given that videos are typically available instead of still images. Temporal pooling of the backbone features over time proves to be effective for the task. Additionally, this approach exhibits faster

inference than alternative techniques. To enhance the reliability of attribute-based person retrieval rankings and address common challenges such as occlusions, an independent hardness predictor is proposed that predicts the difficulty of recognizing attributes in an image. This information is utilized to remarkably improve retrieval results by down-weighting soft biometrics with an increased chance of classification failure. Additionally, three further enhancements to the retrieval process are investigated, including model calibration based on existing literature, a novel attribute-wise error weighting mechanism to balance the attributes' influence on retrieval results, and a new distance measure that relies on the output distributions of the attribute classifier.

Meaningful generalization experiments on pedestrian attribute recognition and attribute-based person retrieval are enabled for the first time. For this purpose, the UPAR dataset is proposed, which contributes 3.3 million binary annotations to harmonize semantic attributes across four existing datasets and introduces two evaluation protocols. Moreover, a new evaluation metric is suggested that is tailored to the task of attribute-based person retrieval. This metric evaluates the overlap between query attributes and the attributes of the retrieved samples to obtain scores that are consistent with the human perception of a person retrieval ranking.

Combining the proposed approaches yields substantial improvements in both pedestrian attribute recognition and attribute-based person retrieval. State-of-the-art performance is achieved concerning both tasks and existing methods from the literature are surpassed. The findings are consistent across both specialization and generalization settings and across the well-established research datasets. Finally, the entire processing pipeline, from video feeds to the resulting retrieval rankings, is outlined. This encompasses a brief discussion on the topic of multi-target multi-camera tracking.

Kurzfassung

Die attributbasierte Personensuche ist eine entscheidende Komponente in verschiedenen realen Anwendungen. Dazu gehören die Videoüberwachung, der Einzelhandel und intelligente Städte. Im Gegensatz zu bildbasierten Ansätzen zur Identifizierung oder Wiedererkennung von Personen zielt sie darauf ab Personen anhand von Beschreibungen ihrer weichen biometrischen Merkmale wie Geschlecht, Alter und Kleidungsfarbe zu suchen. Die attributbasierte Personensuche ermöglicht es beispielsweise Strafverfolgungsbehörden, enorme Mengen an Bildmaterial, das von einem Netzwerk aus Kameras gesammelt wurde, effizient zu durchsuchen, um verdächtige oder vermisste Personen zu finden.

In dieser Dissertation wird ein neuartiger Deep Learning Ansatz für die attributbasierte Personensuche vorgestellt. Das primäre Ziel ist die Erforschung eines ganzheitlichen Ansatzes, der für reale Anwendungen geeignet ist. Daher werden alle hierfür notwendigen Verarbeitungsschritte betrachtet. Die Erkennung von Personenattributen dient in dieser Dissertation als Grundgerüst für die attributbasierte Personensuche. Es werden verschiedene Designmerkmale von Ansätzen für die Personenattributerkennung systematisch auf ihre Eignung für die attributbasierte Personensuche untersucht. Darüber hinaus wird das PARNorm-Modul eingeführt, das die Ausgaben des Attributerkennungsmodells sowohl über die Batch- als auch über die Attributdimensionen normalisiert, um den Einfluss unausgewogener Attribute in den Trainingsdaten zu kompensieren und gleichzeitig die Leistung der Personensuche zu verbessern. Da in der Regel Videos anstelle von Einzelbildern zur Verfügung stehen, werden Strategien zur videobasierten Erkennung von Personenattributen untersucht. Die zeitliche Zusammenführung von abstrakten Merkmalsvektoren, die vom Basismodell generiert werden, erweist sich als effektiv für

diese Aufgabe. Darüber hinaus ermöglicht dieser Ansatz eine schnellere Inferenz im Vergleich zu alternativen Ansätzen. Um die Zuverlässigkeit der attributbasierten Personenerkennung zu verbessern und häufige Probleme wie Verdeckungen zu lösen, wird ein unabhängiger Schwierigkeitsermittler vorgeschlagen, der die Schwierigkeit der Bestimmung von Personenattributen in einem Bild erkennt. Die Schwierigkeitsinformationen werden genutzt, um die Suchergebnisse deutlich zu verbessern, indem schwierige Attribute niedriger gewichtet werden. Zusätzlich werden drei weitere Verbesserungen des Suchprozesses untersucht: eine Modellkalibrierung auf der Grundlage bestehender Literatur, ein neuartiger Mechanismus zur attributweisen Fehlergewichtung, um den Einfluss der Attribute auf die Suchergebnisse auszugleichen, und ein neues Distanzmaß, das auf den Ausgabeverteilungen des Attributklassifikators beruht.

Es werden erstmals aussagekräftige Generalisierungsexperimente zur Erkennung von Personenattributen und zur attributbasierten Personensuche ermöglicht. Zu diesem Zweck wird der UPAR-Datensatz vorgeschlagen, der 3,3 Millionen neue binäre Annotationen verfügbar macht, um die semantischen Attribute von vier bestehenden Datensätzen zu harmonisieren. Zusätzlich werden zwei Auswertungsprotokolle eingeführt. Darüber hinaus wird eine neue Evaluationsmetrik vorgeschlagen, die auf die attributbasierte Personensuche zugeschnitten ist. Diese Metrik wertet die Übereinstimmung zwischen den Attributen der Abfrage und den Attributen der abgerufenen Beispiele aus.

Die Kombination der vorgeschlagenen Ansätze führt zu erheblichen Verbesserungen sowohl bei der Erkennung von Personenattributen als auch bei der attributbasierten Personensuche. Bezüglich beiden Aufgaben wird der aktuelle Stand der Technik übertroffen. Diese Erkenntnisse gelten sowohl für die etablierten Spezialisierungsdatensätze als auch für den UPAR-Generalisierungsdatensatz. Abschließend wird die gesamte Verarbeitungspipeline von den Videodaten der Kameras bis zu den resultierenden Suchergebnissen skizziert. Dies beinhaltet eine kurze Diskussion von Verfahren zum kameraübergreifenden Verfolgen von Personen.

Contents

Abstract . . . . . **i**

Kurzfassung . . . . . **iii**

1 Introduction **1**
1.1 Motivation 1
1.2 Challenges 6
1.2.1 Challenges Emerging from Image Acquisition . . . 6
1.2.2 Soft Biometric-Specific Challenges 9
1.2.3 Real-World Challenges 12
1.3 Contributions 14
1.4 Thesis Outline 16

2 Related Work **17**
2.1 Background on Soft Biometrics 17
2.1.1 Characteristics of Soft Biometrics 19
2.1.2 Soft Biometrics in Crime Investigations 21
2.2 Attribute-Based Person Retrieval 24
2.2.1 Person Retrieval with Natural Language Queries . . 27
2.2.2 Pedestrian Attribute Recognition 28
2.2.3 Learning a Joint Feature Space 37
2.2.4 Summary and Discussion 38
2.3 Further Related Literature 39
2.3.1 Normalization Techniques 39
2.3.2 Hardness and Failure Prediction 41

3 Concept **45**

4 Experimental Setup **53**
4.1 Datasets 53
4.1.1 PETA 54
4.1.2 Market-1501 56
4.1.3 PA-100K 56
4.1.4 RAPv2 57
4.1.5 UPAR 58
4.1.6 MARS 64
4.2 Evaluation Measures 64
4.2.1 Pedestrian Attribute Recognition 65
4.2.2 Person Retrieval 66
4.3 Evaluation Protocols 71
4.3.1 Specialization 71
4.3.2 Generalization 72

5 Baseline **75**
5.1 Problem Formulation 75
5.2 Strong Baseline for Pedestrian Attribute Recognition . . . 76
5.2.1 Backbone 77
5.2.2 Classifier 78
5.2.3 Loss Function 79
5.2.4 Implementation Details 80

6 Pedestrian Attribute Recognition **83**
6.1 Evaluation of Design Choices 84
6.1.1 Binary vs. Multi-Class Attributes 84
6.1.2 Backbone 86
6.1.3 Stochastic Weight Averaging 90
6.1.4 Loss Function 95
6.1.5 Batch Size 99
6.1.6 Dropout 104
6.1.7 Optimizer 106
6.1.8 Label Smoothing 108
6.1.9 Summary 112
6.2 Normalization 115

6.2.1 Methodology . . . 117
6.2.2 Evaluation . . . 119
6.3 Video-Based Pedestrian Attribute Recognition . . . 124
6.3.1 Temporal Pooling . . . 125
6.3.2 Evaluation . . . 127

7 Attribute-Based Person Retrieval . . . 133
7.1 Hardness Prediction . . . 134
7.1.1 Independent Hardness Prediction . . . 136
7.1.2 Self-Referential Hardness Prediction . . . 138
7.1.3 Weight Computation . . . 140
7.1.4 Evaluation . . . 141
7.2 Improvement of the Retrieval Process . . . 151
7.2.1 Reliability Calibration . . . 155
7.2.2 Error Weighting . . . 157
7.2.3 Distribution-Based Distance . . . 158
7.2.4 Evaluation . . . 161

8 Evaluation . . . 169
8.1 Combination of Approaches . . . 169
8.1.1 Specialization . . . 170
8.1.2 Generalization . . . 172
8.1.3 Inference Time . . . 177
8.2 Qualitative Evaluation . . . 179
8.2.1 Specialization . . . 179
8.2.2 Generalization . . . 184
8.3 Comparison with the State-of-the-Art . . . 189
8.3.1 Specialization . . . 189
8.3.2 Generalization . . . 192
8.4 Summary . . . 194

9 Tracking System . . . 197

10 Conclusion and Outlook . . . 207
10.1 Conclusion . . . 207

10.2 Outlook 210

Bibliography **213**

Own Publications **259**

List of Figures **265**

List of Tables **267**

Acronyms **269**

Symbols **273**

1 Introduction

The main objective of this thesis is to examine algorithms for retrieving images or tracks of persons that match a specific semantic description of their visual appearance from a large gallery database collected from a multi-camera network. The focus is on a holistic consideration of the task for real-world applications. This work encompasses all steps required, including the creation of a suitable research dataset and evaluation metric, optimization of Pedestrian Attribute Recognition (PAR) as the feature extraction approach, refinement of the retrieval process, and a brief introduction to the entire system pipeline as well as its implementation.

The following section first provides the motivation for the research in Section 1.1. Subsequently, the most significant challenges hindering robust attribute-based person retrieval are presented in Section 1.2. Finally, the main contributions of the thesis are summarized in Section 1.3 and its structure is presented in Section 1.4.

1.1 Motivation

Rapid advancements in camera technology, combined with a growing demand for security, led to a significant rise of surveillance cameras in both public and private spaces. Nowadays, surveillance cameras are found in various settings, including malls, city centers, airports, and railway stations. Furthermore, such cameras play a crucial role in enhancing the safety of mass events [Sto19]. This trend opened up a multitude of new possibilities, paving the way for smart city applications and enhancing the capabilities of law enforcement agencies. Analyzing the video streams from these cameras enables

intelligent mobility and effective surveillance, all crucial for ensuring public safety during large gatherings or in areas prone to criminal activities [Fen17]. For instance, possible mobility solutions include intelligent routing of crowds or vehicles and traffic light planning [Cor01, Nap22]. Additionally, surveillance camera footage helps in locating missing individuals, keeping track of offenders, and aiding criminal investigations. It empowers authorities to quickly intervene after a security incident and conduct thorough retrograde investigations [Gol23].

Regarding surveillance systems, two primary types of use cases prevail: online and offline evaluation of collected data. Online surveillance systems have been deployed in numerous locations globally, including London [Sat20] in England and Mannheim [dpa20] in Germany. In this context, human operators typically monitor multiple live feeds from areas prone to criminal activities simultaneously, actively seeking anomalies, as illustrated in Figure 1.1a. In the event of a safety-critical incident, online evaluation enables prompt intervention [Akt23]. Moreover, video surveillance systems have played a decisive role in retrograde crime investigations, exemplified by the identification of the Boston bombers in 2013 [BBC13]. In the retrograde use case, law enforcement agencies analyze image and video data collected following an event involving serious criminal offenses. Besides the aforementioned terror attack on the Boston Marathon in 2013, notable incidents include the G20 conference in Hamburg in 2017. The protests during the G20 conference, depicted in Figure 1.1b, escalated into violence and caused extensive crime investigations based on vast amounts of video data [WEL17, Spi18].

However, the amount of data made available grows at impressive rates. Therefore, manual data processing becomes increasingly slow and tedious, demanding a high level of attention. This applies to both use cases. Furthermore, it is challenging to monitor a large number of camera views simultaneously without potentially missing relevant incidents. Automation is essential due to these challenges. When a new suspect is identified, reviewing the video data to assess the suspect's movements and actions becomes a repetitive task. Additionally, ensuring data protection is vital for societal acceptance [Gol22]. Automation alleviates the need for human operators to review entire videos.

Instead, specialized algorithms perform the task efficiently without focusing on specific individuals. Furthermore, automation enables easy enforcement of privacy regulations and accurate logging of data access by human personnel, enhancing overall operational integrity.

(a) Online use case[1]

(b) Retrograde use case[2]

Figure 1.1: Surveillance use cases – Surveillance systems typically serve two distinct use cases. On the one hand, video feeds undergo online evaluation, where data is processed in real-time with the objective of rapid intervention after an incident. As depicted in Figure 1.1a, system operators typically engage in manual efforts to identify relevant activities or persons from numerous concurrently monitored camera streams. In contrast, retrograde evaluation involves the examination of stored data after an incident has occurred. For instance, Figure 1.1b shows scenes from mass protests during the G20 conference in Hamburg in 2017, during which many crimes happened. Although real-time evaluation is no requirement in such scenarios, manual evaluation in the aftermath is a tedious task for human personnel.

Searching for persons of interest in vast amounts of image and video data is among the most critical tasks for law enforcement agencies [Fer14]. Person search allows exploring the occurrences of suspects in the data, tracking their movements, including escape routes, and identifying further crimes associated with the person of interest. Several research fields and definitions exist concerning the use of deep learning techniques for person search, which share commonalities but necessitate distinction due to essential differences [Gal21]. These are person identification, re-identification, and retrieval. Person identification aims at identifying a previously registered individual. This involves

[1] https://www.mannheimer-morgen.de/fotos_fotostrecke,-fotostrecke-modellprojekt-in-mannheim-_mediagalid,32349.html; Christoph Blüthner

[2] https://pixabay.com/de/photos/demonstration-g20-protest-2482170/

acquiring biometric information and comparing it to an identity database. On the other hand, person re-identification focuses on discovering additional occurrences of a human being within the data. Re-identification approaches are applied after an individual of interest is initially observed in the footage, aiming at gathering further information about the suspect's activities. Unlike person identification, there is no direct comparison with an identity database containing biometric data and personal information such as name and address. Furthermore, both hard and soft biometrics may be employed. In contrast to hard biometrics, soft biometrics are semantic features [Sam08, Rei11] that do not allow unambiguous distinction of individuals [Jai04a, Dan16]. Last, person retrieval is a broader task involving finding individuals matching specific characteristics from an image or video database. Contrary to identification, this task does not involve matching to a specific identity, and multiple individuals may be considered correct retrieval results. For this task, soft biometric features are primarily leveraged [Gal21]. For a deeper understanding, the differences between hard and soft biometrics are detailed in Section 2.1.1.

While image-based person re-identification [Ye22] and person identification [Wan22d] are extensively explored research fields, soft biometric person retrieval received comparatively less attention, despite its significance in surveillance applications. In scenarios where no biometric information or clear image of the person of interest are available due to occlusions, low resolution, or camera blind spots, relying on soft biometric descriptions of a person's semantic attributes becomes the only viable option. These descriptions serve as valuable cues for the use as a search query. Typically, these descriptions are obtained from witness testimonies or acquired from the description of relatives, when a missing person is searched.

Semantic person attributes refer to interpretable features with an inherent semantic meaning, for instance, the approximate age and specific clothing types and colors [Dan16]. In the context of this thesis, attributes are treated as discrete entities. This means that search queries contain a list of attributes describing the person who should be found and not, *e.g.*, a description in natural language. The task of finding the occurrences of people based on such information is referred to as attribute-based person retrieval.

Figure 1.2 provides an example of attribute-based person retrieval. The blue box contains the soft biometric query. In this specific case, the list of semantic attributes describes a male senior wearing black trousers and short, white upper-body clothing. Furthermore, the retrieved images should depict individuals wearing a hat. Based on this search query, attribute-based person retrieval generates a ranked list of gallery images. The gallery is composed of person images that should be searched through. Images showing persons who closely match the specified attributes should appear in early positions in the ranking. In contrast, people with vastly different sets of semantic attributes should be relegated to the lower positions of the list. In the example, the top-5 ranking positions in the list are visualized, each identified by a number above the images. The green boxes highlight person images that align with the query description. In contrast, the red boxes denote images depicting individuals whose semantic attributes differ from the query in at least one discrete soft biometric characteristic.

Figure 1.2: Attribute-based person retrieval ranking – The query consists of a list of semantic attributes and is given in the blue box on the left. For instance, a male senior wearing black trousers, short, white upper-body garment, and headwear is searched. The search result is a sorted list of images from a gallery. The numbers above the single images denote the position in the ranking. Red and green boxes illustrate whether the image matches the query (green) or not (red). Source of the images: [Zhe15].

This thesis focuses on developing a deep learning-based framework designed for attribute-based person retrieval in low-resolution, real-world video footage from extensive multi-camera networks. As the base approach, PAR is utilized to extract soft biometric characteristics from image or video

data. The primary objective is to research methodologies that allow seamless integration into smart city or surveillance applications.

1.2 Challenges

While research on related topics such as person re-identification is extensive, performing attribute-based person retrieval in surveillance camera footage remains challenging due to several factors. The challenges stem from differences in camera characteristics, uncontrollable environments, as well as task-specific difficulties. These challenges can be broadly categorized into three groups: challenges related to image acquisition, challenges specific to semantic person attributes, and challenges associated with real-world scenarios. These categories of challenges are discussed in the following sections.

1.2.1 Challenges Emerging from Image Acquisition

First, challenges emerging from the image acquisition process are introduced. This category includes challenging factors originating from the camera sensors, the positions where surveillance cameras are typically installed, and external influences such as lighting conditions. Since this thesis aims at conducting attribute-based person retrieval in data from large multi-camera networks, the approach must be robust against such challenges.

Low spatial resolution: Surveillance cameras are typically strategically positioned to cover the maximum possible area while ensuring protection against tampering. Consequently, the distance to relevant objects is significant, resulting in a spatial resolution of persons being only a few pixels in both height and width. This limited resolution translates to minimal visual information about the appearance of people, especially regarding fine-grained local soft biometrics. Figure 1.3 highlights the complications that low-resolution person crops pose for accurate recognition of semantic attributes by blurring important details.

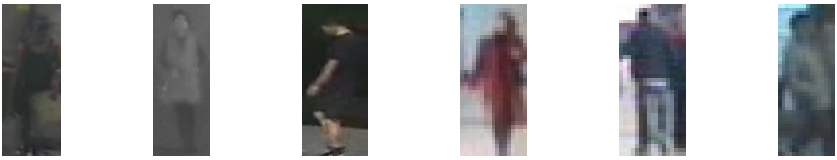


Figure 1.3: Low-resolution imagery – Large distances between surveillance cameras and persons in the background result in low spatial resolution of person images. Only few information about the visual appearance is included so that the recognition of attributes is greatly impeded. Image sources: [Den14, Liu17].

Viewpoints: Environments in realistic scenarios are typically uncontrollable, resulting in images or videos being recorded from various cameras with differing viewpoints and viewing angles. Cameras in video surveillance networks are mounted at different heights and with different steepness in dependence on the area they oversee. For instance, high-mounted overview cameras monitor public squares, while low-mounted cameras are applied indoors or in narrow streets. Furthermore, additional data recorded by witnesses with their smartphones or by private surveillance cameras might be considered. As a results, persons appear in different poses, and small-scale attributes, *e.g.*, glasses, may not be determinable depending on the pose. Moreover, attributes such as the appearance of a backpack vary strongly depending on the viewing angle. For instance, the entire backpack is visible from the back, whereas only the straps are visible in a frontal view of a person. This challenge is demonstrated in Figure 1.4, where the first four images starting from the left depict the same person from different viewing angles. The subsequent images illustrate the varying appearance of a backpack depending on whether the person wearing it is seen from the front or back.

Figure 1.4: Different viewpoints – This figure shows samples for two persons captured from different point of views. The first four images depict the person from a steep perspective and viewed from the right, front, or left side. Especially, recognizing attributes such as backpack suffers from varying viewpoints due to different appearance when seen from the front or back. Image sources: [Den14, Zhe15].

Illumination: Illumination greatly impacts image or video quality. Inadequate illumination, such as during the night or in shaded regions, results in low contrast, blurring, and raised image noise. Additionally, it leads to poor color perception if the illumination is too low. In contrast, excessive illumination for the camera sensor causes details to be lost in overexposed images. Moreover, varying lighting conditions may coexist in the same scene. For instance, the visual appearance changes when a person walks into a shaded area. Figure 1.5 presents a comparison of images from two individuals captured under different lighting conditions. Dependent on the illumination, especially colors appear different, and details are hardly visible. Note that the person on the right wears the same clothes in all images.

Figure 1.5: Illumination – Two persons with three images each that show the difficulty with illumination changes. For instance, colors appear strongly different depending on the current lighting conditions. Image sources: [Den14, Zhe15].

Image noise: Image noise refers to the degradation of captured pixel values in an image caused by random disturbances or inconsistencies. Noise mainly originates from the sensor, loss of information during quantization, and statistical photon effects [Bov05].

Blurring: Blurring in images or specific image regions results in decreased sharpness and weak contrast. It occurs due to moving objects, camera motion, or image regions that are out of focus. Blurred images are challenging as fine-grained information remains vague.

Indoor and outdoor scenes: Surveillance cameras are deployed both indoors as well as outdoors, resulting in varying lighting conditions. For instance, outdoor imagery is influenced by changing weather conditions, whereas malls, airports, and other public spaces typically have uniform artificial lighting indoors. This contrast can be observed in Figure 1.6, where

the first three images from the left were recorded indoors and the subsequent ones were collected outdoors.

Figure 1.6: Indoor and outdoor scenes – The first three images from the left display images captured indoors, while the following images were taken outdoors. The differences in illumination are clearly noticeable. Indoor scenes benefit from consistent artificial lighting, resulting in well-illuminated and comparable conditions. On the other hand, outdoor illumination is greatly affected by various factors and frequently undergoes changes. Image sources: [Den14, Li19a].

Camera types: Various types of cameras possess distinct characteristics, including differences in internal calibration and sensor type. Consequently, images may exhibit distortions or variations in colors across different camera models.

In summary, several significant challenges are related to the image acquisition process. These challenges are not specific to a particular task but rather present difficulties for most computer vision tasks applied to real-world imagery.

1.2.2 Soft Biometric-Specific Challenges

Contrary to the challenges discussed before, the challenges presented in this section stem from the recognition of soft biometrics or similar imbalanced, fine-grained classification tasks. Soft biometrics are semantic features of individuals describing an individual in a humanly understandable manner [Dan16].

Occlusions: Occlusions pose a substantial challenge since relevant characteristics may not be visible in the person image. For instance, if a person's lower-body is occluded, reliable recognition of related soft biometrics becomes impossible. This challenge is visualized in Figure 1.7. Occlusions can result from

various sources, including objects, other persons, or even the depicted person's own body parts. When it comes to attribute-based person retrieval, occlusions pose a significant challenge that must be effectively addressed to achieve accurate retrieval results.

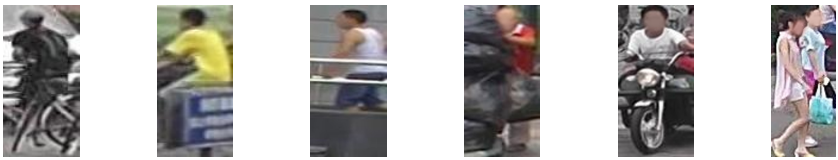


Figure 1.7: Occlusions – Occlusions caused by objects, person, or obstacles are a severe challenge in PAR and attribute-based person retrieval. Relevant characteristics might not be visible and, therefore, attributes not determinable. Image sources: [Zhe15, Liu17].

Intra-class variance: Semantic attributes exhibit a wide range of variance in their appearance. Colors, for instance, can hardly be assigned to distinct classes due to the fluid transitions between them. Furthermore, attributes like attachments occur in various colors and shapes, requiring a deep learning model to abstract and focus on relevant characteristics in order to be able to recognize these attributes robustly. This challenge exacerbates in cases of generalization, when a deep learning model is transferred to a new domain with different characteristics. Examples illustrating this variance are provided in Figure 1.8. Hats and handbags, for instance, occur in a wide range of shapes. Similarly, colors appear in different shades.

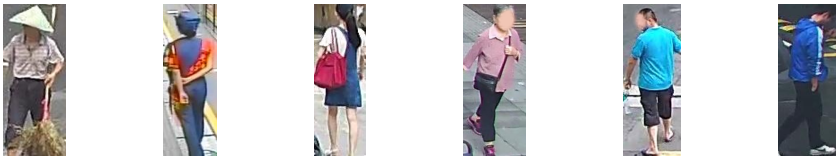


Figure 1.8: Intra-class variance – The selected images showcase large intra-class variance. Hats and handbags appear in different shapes, colors, and sizes. Similarly, distinct colors such as blue occur in a large variety of shades. Image sources: [Liu17].

Imbalanced data distributions: As shown in Figure 1.9, imbalanced attribute distributions present a further challenge.

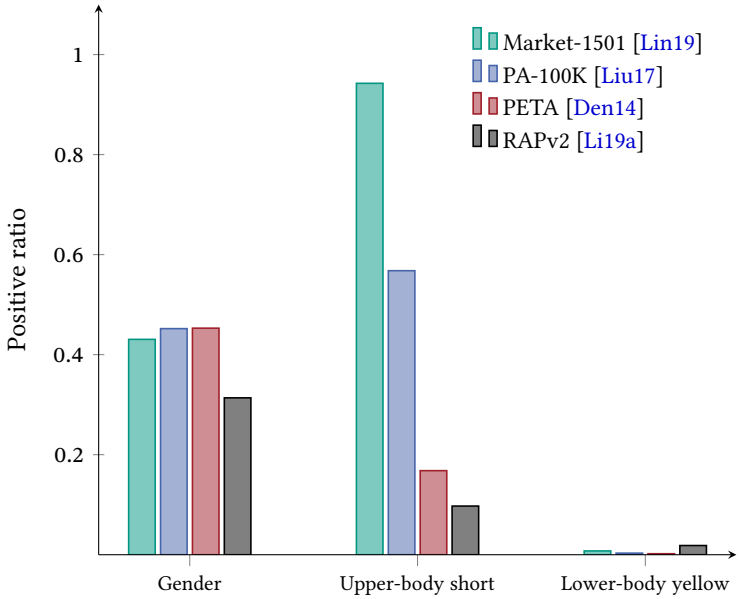


Figure 1.9: Attribute distributions – Positive ratios of selected attributes are reported for different datasets. This refers to the ratio of samples that exhibit a particular attribute. One can observe that attributes such as gender are almost equally distributed on each of the datasets. However, for attributes such as upper-body clothing length that are highly dependent on the season, severe distribution shifts across the datasets are visible. Also, there are rarely occurring attributes (*e.g.*, lower-body clothing color yellow) that are therefore difficult to learn from only a few samples.

While attributes such as gender are almost equally distributed, there are only a few samples for certain attributes, like specific colors. These imbalances arise from both real-world attribute distributions and dataset biases. Machine learning models trained on imbalanced attribute distributions tend to prefer recognizing either the presence or absence of the attributes, depending on their distribution in the training set. Consequently, these models generalize poorly to unseen data with deviating attribute distributions. Notably, differences regarding attribute distributions are observable across several datasets, particularly apparent for attributes such as upper-body clothing length.

Biased and limited datasets: The last soft biometric-specific challenge stems from the unavailability of appropriate datasets. Available surveillance datasets [Zhe15, Den14, Li19a, Liu17] are comparatively small and include huge biases. For instance, the Market-1501 dataset was captured outdoors on a campus during summer, featuring predominantly young people wearing short clothes. Furthermore, appearance of persons but also clothing and accessories may be specific to the culture and country in which the data was collected. Consequently, it becomes challenging to avoid deep learning models from overfitting training data, which is essential for achieving strong generalization capabilities across domains. Moreover, each dataset comes with its own set of attribute annotations, making it difficult to conduct cross-domain experiments to compare the generalization ability of deep learning models.

In conclusion, recognizing soft biometric attributes in images, as the basis for attribute-based person retrieval, introduces task-specific challenges. Among these challenges, occlusions and imbalanced attribute distributions within datasets stand out as the most significant. Additionally, there is a notable lack of suitable datasets for adequately assessing and comparing the generalization capabilities of deep learning methods in this context.

1.2.3 Real-World Challenges

The final category of challenges to consider are real-world challenges that arise from the deployment of deep learning methods in practical applications.

Generalization ability: Machine learning models are trained on small excerpts of reality, which are often biased in terms of data distribution and diversity. However, robustness against variations in data is essential to ensure optimal performance on previously unseen domains. For instance, deep learning models should capture general concepts behind semantic attributes rather than focusing solely on the exact samples from the training data.

Video processing: While most research datasets are image-based, *i.e.*, attribute recognition and person retrieval are carried out on single images, primarily videos are processed in a surveillance scenario. Hence, it is crucial to research methods that recognize soft biometrics based on whole tracklets of people and not individual images.

Processing pipeline: In real-world scenarios, persons need to be detected and tracked in videos before attribute-based person retrieval is performed. Therefore, an entire pipeline consisting of grabbing the camera feeds, detecting persons in each video frame, and tracking the people over time is necessary. Errors in pipeline components, such as misaligned bounding boxes generated by the person detector or incorrect tracks, present additional challenges for attribute-based person retrieval.

Real-time computation: Meeting real-time processing requirements is essential for covering both the offline and online use case. Rapid intervention after an incident necessitates searching for persons in current video frames without delay. Thus, applied deep learning models need to be computationally efficient.

Hardware: Given the typically high number of cameras, substantial computing resources, especially Graphics Processing Units (GPUs), are necessary. Hence, an attribute-based person retrieval system in its entirety is required to run on affordable hardware, considering that financial resources are often limited.

In summary, the real-world application of attribute-based person retrieval approaches necessitates additional processing steps as well as video processing. These steps must adhere to real-time requirements to ensure applicability in the online use cases, where incoming videos from the cameras are processed in real-time, enabling quick intervention after security-relevant incidents.

1.3 Contributions

This thesis aims at designing a deep learning-based framework for attribute-based person retrieval in multi-camera networks under realistic, real-world settings. The main contributions are summarized as follows:

- The utilization of PAR as the basis for attribute-based person retrieval is extensively analyzed. Prior studies primarily concentrated on optimizing the PAR task, disregarding the specific challenges and requirements originating from the retrieval task. In this thesis, the impact of design choices in the development of PAR methods are thoroughly examined, focusing on crucial decisions to improve both PAR and attribute-based person retrieval accuracy simultaneously [Sch18, Spe20b, Cor23, Spe23b]. Additionally, a normalization module is proposed to specifically address the retrieval-relevant issue of imbalanced PAR model outputs, enhancing attribute recognition and attribute-based person retrieval [Spe23a].
- This thesis evaluates recent strategies for video-based PAR, with the aim of aggregating information across tracks to enhance the robustness of attribute-based person retrieval [Spe20c]. The findings demonstrate that the proposed lightweight temporal pooling model is superior to more complex approaches in PAR and attribute-based person retrieval.
- Explicit difficulty prediction for multi-label classification is introduced to identify indeterminable semantic attributes, *e.g.*, due to occlusions or similar challenges [Spe20a]. The proposed Hardness Predictor (HP) provides important complementary information, surpassing uncertainty estimation methods when utilized as a weighting mechanism during person retrieval [Flo21].
- The outputs of a PAR model need further processing to enable effective attribute-based person retrieval. Specifically, this involves computing distances between queries and gallery samples to determine the similarity. This thesis delves into this aspect and

proposes several measures to enhance the reliability and quality of the resulting attribute-based person retrieval rank lists [Spe21a]. Reliability calibration is applied to compensate for over- or underconfidence of the attribute classifier. Additionally, a weighting technique is introduced to balance the impact of the attributes on the retrieval distance, alongside a novel distance metric that takes into account the output distributions of the classification model.

- This thesis introduces the Unified Pedestrian Attribute Recognition (UPAR) dataset, a pioneering dataset enabling large-scale generalization experiments concerning PAR and attribute-based person retrieval [Spe23b]. The dataset unites image data from well-known PAR datasets and harmonizes annotations for 40 attributes by contributing 3.3 million new binary annotations. Comprehensive studies focused on generalization are conducted, with the findings integrated to develop more robust models [Cor23, Spe23b].
- Existing person retrieval metrics fall short in adequately measuring the usefulness and meaningfulness of rank lists to the system operator for attribute-based person retrieval. This is due to the practice of making binary relevance decisions, even though certain attributes might be occluded, and the image might indeed depict the person of interest. To address this limitation, a novel measure is contributed that considers the degree of match between the query attributes and the soft biometrics annotated for the gallery samples [Spe23a].
- An extensive evaluation conducted on both publicly available benchmarks and the proposed UPAR dataset demonstrates the superior performance of the proposed framework compared to existing literature. The results establish a new state-of-the-art for both PAR and attribute-based person retrieval tasks in various settings, including specialization and generalization. This achievement proves the transferability and practical utility of the proposed methodology in real-world applications.
- A brief overview of further works by the author is provided, demonstrating the integration of the proposed attribute-based person

retrieval approach into a surveillance system that covers the entire processing pipeline, including person tracking, and achieves real-time computation [Köh20, Spe22a, Spe22c].

1.4 Thesis Outline

This thesis is structured as follows. First, related literature is reviewed in Chapter 2. This includes background information on soft biometrics, the general presentation of works existing in attribute-based person retrieval and PAR, and closely related works w.r.t. the proposed methodologies. Subsequently, Chapter 3 introduces the overall concept that is followed in this thesis. In Chapter 4, the datasets and evaluation protocols are presented, including the UPAR dataset and the novel evaluation measure. Chapter 5 contains the description of the baseline method. The following Chapter 6 deals with improvements in terms of the PAR model that serves as the feature extraction approach for attribute-based person retrieval. Methods regarding optimal design choices, normalization, and video-based processing are presented and evaluated in detail. In Chapter 7, approaches concerning the retrieval process are focused. Specifically, the impact of considering indeterminable attributes, calibrating the PAR model's outputs, and weighted distance computation are investigated. The combination of the proposed methods in this thesis is evaluated in Chapter 8. Furthermore, results are compared to the current state-of-the-art to prove the effectiveness. Next, an entire surveillance system is introduced in Chapter 9 in which the attribute-based person retrieval is embedded to enable the search for persons in multi-camera networks. This includes Multi-Target Multi-Camera Tracking (MTMCT) as well as an efficient implementation in order to allow real-time processing of video feeds. Finally, Chapter 10 sums up this thesis and provides possible research directions for future works.

2 Related Work

This thesis aims at researching attribute-based retrieval of persons using deep learning techniques. This chapter presents relevant literature related to this topic and the proposed framework.

First, Section 2.1 provides background information on soft biometrics and their use in crime investigations. Next, the literature on attribute-based person retrieval is presented in Section 2.2, where possible approaches to the task are delimited. Last, additional research fields relevant to specific optimizations proposed in this thesis are concisely introduced in Section 2.3.

2.1 Background on Soft Biometrics

In this thesis, soft biometric features, often referred to as semantic attributes, serve as queries for conducting person retrieval. To provide a comprehensive understanding and background on soft biometrics, this section first delves into defining soft biometrics and distinguishing them from hard biometrics. Subsequently, characteristics of soft biometric information are introduced in Section 2.1.1, followed by a brief excursus on the use and significance of such features in crime investigations in Section 2.1.2.

According to Jain et al. [Jai08], biometrics are physical, behavioral, or physiological attributes unique to a specific individual, such as fingerprint, face, iris, gait, or voice. These attributes, often termed hard biometrics, allow the unequivocal identification of human beings. In contrast, soft biometric features cannot be assigned to a specific individual unambiguously. Multiple individuals may share the same set of soft biometric attributes, making unique identification impossible.

The utilization of soft biometric features for person identification in law enforcement applications dates back to Alphonse Bertillon [HTF56] in 1956. Bertillon proposed leveraging anatomical, morphological, and anthropometrical characteristics, in addition to profile and full-face images, to identify criminals. In related literature, there are several definitions of the term soft biometrics [Jai04a, Sam08, Rei11, Dan16], which are summarized as follows: Soft biometrics comprise characteristics that humans typically use to describe each other [Sam08, Rei11]. These features aid in recognizing individuals without allowing for clear-cut distinctions between people [Jai04a, Dan16].

Examples of soft biometric person attributes are given in Figure 2.1. The figure shows attributes and their classification according to the taxonomy proposed by Dantcheva et al. [Dan16]. The taxonomy distinguishes four categories: demographic, anthropometric, medical, and material and behavioral attributes. Demographic attributes are related to a person's inherent characteristics, including age, gender, ethnicity, and physical traits such as the color of eyes, hair, and skin. These attributes play a significant role in person identification and retrieval since they can be disguised but not completely altered. Anthropometric soft biometric features comprise body measurements like body height and proportions of facial features. Although being frequently reported by eyewitnesses, this information is often subject to uncertainty and, thus, is challenging to use for person retrieval in real-world scenarios [Spo92, Fli86]. More details concerning the utilization of soft biometrics in crime investigations are provided in Section 2.1.2. The third category relates to medical properties. Attributes within this category are difficult to capture at a glance by eyewitnesses and, thus, are of minor importance in the context of person retrieval. Attributes concerning objects and accessories a human may carry belong to the material and behavioral soft biometrics category. For identification tasks, these characteristics are not reliable, given that they can be easily changed. However, in the context of person retrieval within a surveillance scenario, they represent crucial information for finding a person of interest during a limited time frame [Jai08, Dan11, Dan16].

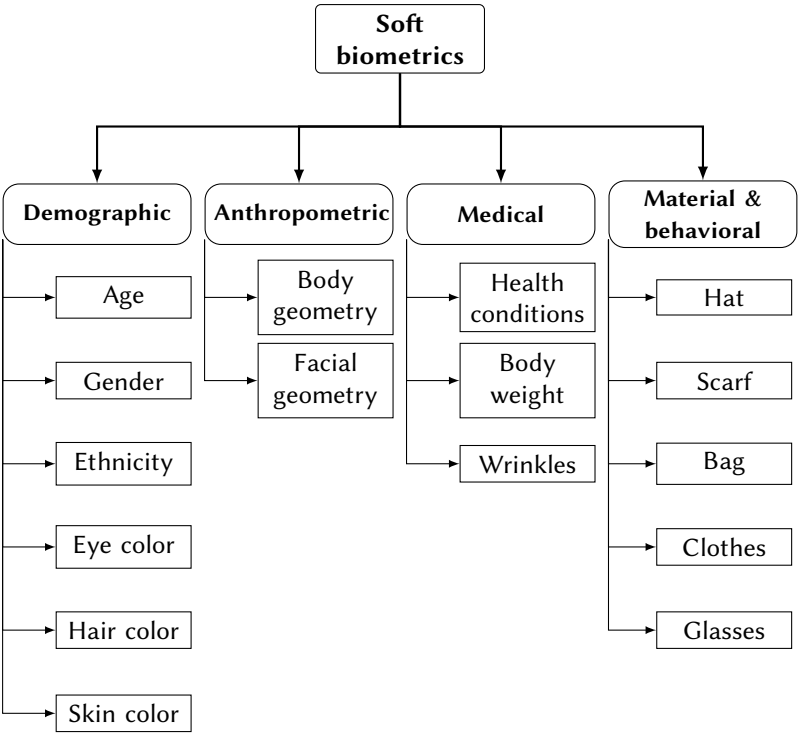


Figure 2.1: Taxonomy of soft biometrics – Following Dantcheva et al. [Dan16], soft biometrics are categorized into four groups: demographic, anthropometric, medical, and material and behavioral attributes. Example characteristics for each category are given. Changed representation after [Dan16].

2.1.1 Characteristics of Soft Biometrics

After providing a definition of soft biometrics in the preceding section, the important characteristics of this feature type are detailed. Hard biometric features come with serious drawbacks in comparison to soft biometrics. They are primarily used to identify individuals, which poses a direct threat to privacy and may lead to severe consequences for incorrectly identified individuals. For instance, the city of London implemented a face recognition system into

its existing surveillance camera network. However, despite achieving high accuracy scores for this task on benchmark datasets, human operators observed poor outcomes in the real-world scenario [Sat20]. Independent studies found that the identification accuracy was below 20% during the time period under consideration. Challenges such as unfavorable viewing conditions in uncontrollable environments made it exceedingly difficult to acquire sufficiently accurate information from the video data. This section draws a comparison between hard and soft biometrics, elaborating on the essential characteristics of soft biometrics.

Privacy-preserving: Soft biometric attributes are not unique to a specific individual, making them considerably more privacy-friendly than hard biometric features [Jai04a, Jai04b, Dan16]. Multiple persons may share the same attributes and fit the same attribute description. In addition, certain attributes, particularly those related to material soft biometrics, change regularly, allowing only short-term re-identification. For instance, persons of interest can only be re-identified if their clothing remains unchanged. However, this characteristic also presents a disadvantage as criminals may exploit this. Moreover, the absence of a requirement for registration of individuals and an identity database is a notable advantage. This not only prevents the capture and theft of sensitive data, *e.g.*, by hackers, but also upholds privacy and data security.

Low hardware requirements: Soft biometrics are non-invasive methods, allowing for easy extraction of such characteristics. No dedicated sensor hardware such as fingerprint or iris scanners is required. Inexpensive cameras are sufficient to acquire data to recognize semantic attributes from a distance [Jai04b].

No cooperation required: Another advantage is that no subject cooperation is required [Jai04b], and attributes are determinable after an incident. Even criminals attempting to prevent being captured by biometric systems are captured since it is difficult to escape or avoid cameras entirely, *e.g.*, during mass events or in public transportation [Fen17].

Recognizable at distance: Due to the acquisition of data through simple cameras, semantic attributes are extractable from a distance without requiring interaction with the subject [Jai04b]. Furthermore, semantic features such as colors can be determined even in low-resolution imagery and are not significantly affected by the viewing angle. In contrast, camera-based hard biometrics, like face identification, necessitate frontal views of the face with high resolution to enable accurate identification. However, in an uncontrollable real-world scenario, fulfilling this requirement is often challenging [Sat20].

Semantic meaning: An increasingly important factor is the semantic meaning of soft biometric person attributes. According to the definitions [Jai04b, Sam08, Rei11, Dan16], soft biometric characteristics align with the attributes people naturally use to describe one another. The semantic meaning contributes to understanding how the retrieval result was obtained. Providing the recognized semantic attributes for each gallery image, along with the degree of fit to the query for each attribute, provides meaningful additional information to the system operator. This information assists in identifying latent biases or data-specific issues. Consequently, it allows for adapting the query accordingly to improve the search results or providing feedback to the system developer for enhancement. In contrast, typical image-based identification or re-identification approaches generate abstract global or local feature vectors that lack semantic meaning and human comprehensibility [Wan22d, Ye22].

2.1.2 Soft Biometrics in Crime Investigations

Soft biometric features in crime investigations mostly appear as person descriptions obtained by eyewitnesses' testimonies, which in turn are essential for investigators in many crimes. Such descriptions may be used for locating a person of interest after a security-relevant incident but also serve as the basis for identifying suspects from mug shots, creating phantom images, or choosing filler persons to arrange lineups [Mei07]. The first application corresponds to the attribute-based person retrieval task investigated in this thesis. Person descriptions constitute attribute-based person retrieval queries to locate suspects within image or video data. Hence, this section sheds light

on person descriptions in the context of law enforcement and discusses the implications for the development of algorithms for attribute-based person retrieval. In general, related research distinguishes studies based on real crimes and fictional studies conducted in laboratories. However, results vary significantly between both categories [Mei07], therefore, this section focuses on real-world studies.

Number of features: The number of features mentioned by eyewitnesses is of great importance for narrowing down the search space of matching individuals. Sporer [Spo92] found that person descriptions are often vague and not specific. Witnesses in their study reported 9.71 ± 7.03 different characteristics in their descriptions. The high standard deviation is particularly striking. Many people described the suspects with only a few soft biometrics. Generally, the number ranged from 1 to 48 different details. These findings are consistent with those of Kuehn [Kue74] and Lindsay et al. [Lin94]. Kuehn [Kue74] found that a majority of eyewitnesses declare either 8 or 9 semantic attributes. Lindsay et al. [Lin94] observed that with 3.94 features on average, even fewer features were reported. In comparison, according to the work of Lindsay et al. [Lin94], person descriptions obtained during laboratory experiments contained 7.35 characteristics on average. Explanations for this finding might include the weapon-focus effect [Ste92] or increased stress levels in real crime situations. Such influential factors are discussed in more detail later in this section. In terms of attribute-based person retrieval, these findings indicate that it is essential to have multiple witnesses and, thus, person descriptions to obtain a complete description of the person of interest. As a result, the methodology should be able to process merged testimonies.

Most frequent descriptors: When considering the descriptors most frequently mentioned by witnesses, it is noticeable that mainly general information [Kue74, Lin94] and statements concerning clothing are made [Spo92, Lin94]. For instance, Kuehn [Kue74] examined eyewitness testimonies collected immediately after an incident, and particularly general descriptions of the offenders were given. The most frequently mentioned attributes include gender, age, and height. In this study, only a tiny fraction of witnesses

recalled fine-grained information such as eye or hair color. Slightly different observations were made by Sporer [Spo92]. In the archival analysis of real crime testimonies, most frequently named descriptors referred to clothing followed by information about the head or face. Facial features in this case mainly contain the hair color of people. Descriptive details about the general impression of a perpetrator were provided only third most often but still accounted for 25% of the descriptors. Lindsay et al. [Lin94] compared the most likely stated attributes by witnesses of staged crimes with those reported by real crime witnesses. Interestingly, participants in the laboratory studies often mentioned more specific information like clothing or hair color but forgot general information such as gender (less than 50% of the participants). Contrary to this, especially gender (96%) was mentioned in the vast majority of real-world cases. The second most frequently reported were characteristics related to clothing, followed by hair color. Similar to the studies above, fine-granular facial features (10%) were mentioned by only a tiny proportion of witnesses, possibly because these features are difficult to perceive during a crime. In conclusion, eyewitnesses of real crimes often report general impressions and clothing information which are, therefore, essential characteristics to consider during attribute-based retrieval. Highly-localized attributes, particularly facial features, are seldom included in person descriptions due to the difficulty of perception in stressful situations or from a distance.

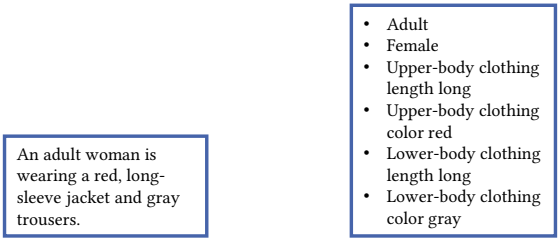
Accuracy of descriptions: In general, the accuracy of person descriptions acquired from witness statements is mostly good according to relevant studies [Yui86, van97]. Both works observed person descriptors more likely to be correct than faulty. Yuille et al. [Yui86] specified the accuracy across all kinds of descriptors with 76% for the investigated shooting incident. Even after four to five months, witnesses could provide highly accurate descriptions when they were interviewed a second time. However, it was found that person descriptions are less precise than object or action descriptions in the same shooting case. The reliability of descriptors varies greatly depending on their type. For instance, estimates of age, height, and weight are prone to errors (23% error rate according to [Yui86]) due to various factors. First, height and weight estimates are primarily specified as average by witnesses, which leads

to an effect called regression to the mean [Spo92, Fli86]. For instance, eyewitnesses tend to underestimate the size of tall people and overestimate the height of short people, respectively. Besides inaccuracies, this finding negatively impacts the usefulness of such features for person search since these attributes are not distinctive. Second, the own-anchor effect refers to witnesses comparing the perpetrators to themselves [Fli86]. Thus, the descriptions of persons concerning characteristics such as body size or weight cannot be assumed to be objective. Details about color and style of clothes and hair were erroneous to 18% in the study of Yuille et al. [Yui86]. Issues regarding the memory of colors of human beings may explain these errors [Mün15]. The work of van Koppen et al. [van97] indicates huge error rates related to facial features, such as eye color, mouth, or nose. More than 60% of the witness statements concerning these characteristics were erroneous. Local features are rarely mentioned and are also subject to a high degree of uncertainty. Besides, the study indicates a negative correlation between the accuracy of a person description and its completeness, *i.e.*, the more extensive a description, the less accurate.

2.2 Attribute-Based Person Retrieval

In general, person retrieval can be divided into two research fields based on the type of the query: image-based or text-based. This thesis addresses person retrieval based on textual queries describing soft biometric attributes. In this case, no probe images of a person of interest are required to find matching individuals in image or video databases. Instead, semantic person descriptions serve as input to the retrieval system. These soft biometric retrieval techniques are further partitionable based on the form of the textual queries [Gal21]. Figure 2.2 highlights the differences between natural language queries, which describe a person's visual appearance using whole sentences, and discrete attribute queries that only include a list of certain semantic attributes. In the context of surveillance applications, there are multiple reasons why discrete attribute queries represent the more important form. For instance, testimonies are subject to uncertainty, as discussed in Section 2.1.2.

Hence, system operators may start searches with only parts of the information witnesses provide or combine statements of several witnesses, which is easier in the discrete form. Furthermore, additional complexity and error sources by natural language processing are avoided. Although this thesis focuses on the discrete form, Section 2.2.1 offers a concise summary of works related to natural language-based person retrieval in order to provide a complete picture.



(a) Natural language **(b)** Discrete attributes

Figure 2.2: Types of textual queries – Attribute-based person retrieval systems are dividable according to the accepted input. Textual queries can either be in the form of sentences expressed in natural language or in a discrete manner.

Whereas natural-language-based approaches are typically restricted to learning a shared cross-modal feature space between image and text embeddings, discrete attribute-based person retrieval allows for a wider range of approaches. Methods can be categorized into three distinct types based on the feature level at which the alignment between discrete textual descriptions and images takes place, as illustrated in Figure 2.3.

The first approach involves employing a PAR model to determine the semantic attributes of people for the images in the gallery. Once identified, these attributes are compared to the attribute query to evaluate the similarity. This method preserves semantic information and, therefore, allows to understand and interpret the retrieval results by analyzing the attributes predicted for the images. In addition, the semantic information extracted is beneficial to other applications, for instance, as complementary data to improve image-based person re-identification or tracking [Zhu15, Lin19, Spe20b]. For these reasons,

the PAR approach is preferred in this thesis to tackle the task of attribute-based person retrieval. The related literature regarding PAR is presented in Section 2.2.2.

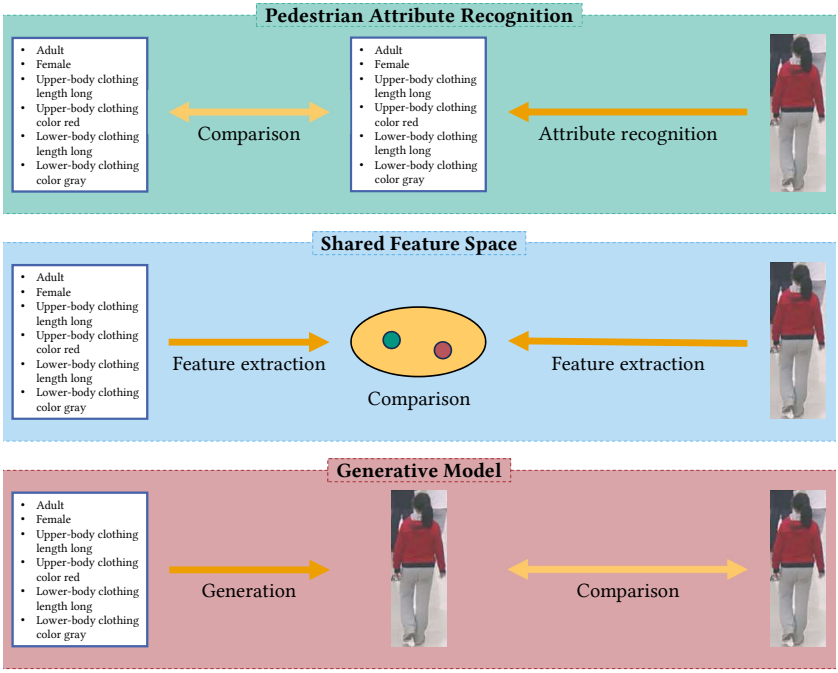


Figure 2.3: Approaches for person retrieval using discrete attribute queries – There are three approaches to address this task, which differ based on where the comparison between text and image modalities is performed. The first approach uses PAR to determine the semantic attributes of the gallery images, which are then compared to the query attributes. The second approach involves learning a joint feature space between both modalities, where the inputs from both text and image modalities are embedded. Finally, attribute queries can be used to generate an image of the described person, which is then compared with gallery samples in the image space.

The second approach aligns discrete attribute queries and images in a shared feature space. This necessitates the implementation of feature extractors for

both modalities, similar to natural language-based techniques. Typically, binary vectors are utilized to encode discrete textual queries, where each element represents a semantic attribute, and its value indicates whether it is present or absent. These vectors serve as the input for the text encoder. The rankings for retrieval are generated by calculating the distance within the shared feature space between the attribute embeddings and image embeddings, which are created by the image encoder. Research on learning a joint feature space between images and discrete attribute queries is outlined in Section 2.2.3.

Last, it is possible to compare textual queries and images at the image level. For this purpose, images of persons must be created on the basis of the attribute description provided in the query. Either pixel-based comparison or a retrieval based on image-to-image person re-identification models may be performed to assess the similarity and build the ranking. However, due to the high complexity and number of degrees of freedom involved in generative models, this method is not commonly used in the related literature. Thus, option three is not covered in the following literature review.

2.2.1 Person Retrieval with Natural Language Queries

Only a few publicly available datasets provide surveillance imagery and descriptions in natural language of the depicted individuals. Researchers focus predominantly on the CUHK-PEDES [Li17b] dataset, which comprises 40,206 images illustrating 13,003 unique identities. This dataset features an impressive set of 80,412 textual descriptions at the identity level. In comparison, the ICFG-PEDES dataset [Din21] includes solely 54,522 image-level annotations of natural language descriptions.

Natural language-based person retrieval approaches aim at aligning text and image features in a joint feature space, using two encoders that generate representations for the image and text input, respectively. Early works [Li17a, Li17b, Che18] leverage VGG [Sim14] and Long Short-Term Memory (LSTM) [Hoc97] for extracting image embeddings and textual features, respectively, and apply matching losses for alignment. With the

progressive development of enhanced feature extractors, encoder backbones for both image and text embeddings were replaced with modern variants. Primarily ResNet [He16] derivatives are applied by later works to extract image features [Sar19, Niu20, Zhe20, Che21] but also Convolutional Neural Networks (CNNs) optimized for the use on mobile devices [Zha18a, Wan19b, Agg20], such as MobileNet [How17]. Concerning the extraction of textual features, the focus of literature [Sar19, Wu21, Zhu21, Che22a, Li22, Sha22, Wan22e, Wan22f, Shu23] shifted to BERT [Dev19], *i.e.*, transformer-based architectures since they established themselves as state-of-the-art techniques in the natural language processing domain.

Developing novel loss functions to improve cross-modal alignment of image and text features is a vital area of research [Li17a, Zha18a, Sar19, Wan19b, Zhe20, Che21]. Further research directions are similar to methods employed in related topics such as image-based person re-identification or PAR. For instance, approaches exploit additional branches that encode local information about person segmentation [Wan20], human body parts [Che22a], or colors [Wu21, Wan22e]. Moreover, further works [Din21, Sha22] rely on implicit attention mechanisms to improve the localization of relevant features. The primary disadvantage of such approaches is the increased complexity due to additional branches and, thus, worse inference times.

All previously mentioned methods utilize data from only one modality for their backbone model pre-training. However, the emergence of CLIP [Rad21] has enabled several works [Han21, Yan22, Jia23] to pre-train their models with image-text pairs in order to improve retrieval accuracy.

2.2.2 Pedestrian Attribute Recognition

Recognizing semantic attributes of individuals, commonly referred to as PAR, is a challenging task in computer vision. This section provides an overview of the related research in this area. Note that traditional methods using handcrafted features are not covered, as deep learning models offer improved performance by automatically learning complex representations from image data [Li15].

The methods are categorized into global image-based models, part-based approaches, attention mechanisms, novel loss functions, and transformer-based methods. This categorization orients itself on the survey of [Wan22c]. Further directions of research concerning PAR include sequential predictions of attributes [Liu18a, Zha19b], curriculum learning approaches [Sar18a, Wan19a], graphs [Li19b, Tan20, Lu23], and reinforcement learning [Ji22]. However, since these methods are of minor importance to this work, they are not covered in the following analysis. More details are given in the comprehensive survey by Wang et al. [Wan22c]. Finally, methods concerning video-based PAR are presented.

Global image-based methods: Global image-based models process the entire input image as a whole, rather than dividing the image into several subregions. Consequently, these models are typically computationally efficient, which makes them suitable for real-world deployment. The ACN model by Sudowe et al. [Sud15] utilizes a pre-trained AlexNet as the backbone and employs distinct loss function for each attribute. The loss functions rely on the Kullback-Leibler divergence between the model's attribute predictions and ground truth labels. The DeepSAR and DeepMAR models [Li15] mark a significant milestone in deep learning-based PAR. Besides showcasing substantial improvement compared to traditional methods, the authors demonstrate that jointly learning multiple attributes within the same neural network (DeepMAR) is superior to attribute-specific models (DeepSAR). This is attributed to the implicit learning of attribute correlations, such as co-occurrences. DeepSAR utilizes the softmax cross-entropy loss function and two output neurons, with one for each attribute's presence or absence. In contrast, DeepMAR relies on the sigmoid cross-entropy loss function for multi-label classification. Furthermore, the issue of imbalanced label distributions, which is commonly encountered in real-world scenarios, is addressed through the introduction of a loss weighting function. This function assigns a weight to the attribute-specific loss value based on the ratio of its occurrence in the training set, also known as the positive ratio of attributes. Training samples depicting a rare attribute receive a greater weight than samples lacking this attribute, effectively guiding the model's focus toward those samples to enhance the recognition of attributes with limited instances. Contrary to

DeepMAR, which implicitly learns the correlations between attributes, Han et al. [Han19] explicitly exploit priors about the co-occurrence of semantic attributes. Their CoCNN model features a multi-branch architecture that generates separate attribute predictions in each branch. The predictions are adjusted based on the co-occurrence priors and fused with the original predictions with an attribute-aware pooling method. However, according to Zhou et al. [Zho23], the generalization capability may be negatively impacted by co-occurrence biases as inter-attribute dependencies are specific to the dataset. Additionally, limited datasets may result in learning unwanted correlations between attributes, leading to recognition errors when inferring attribute predictions based on unrelated attributes. The decoupling of attributes is accomplished in this study by reducing the mutual information in the features among attributes. Jia et al. [Jia21b] revisit the task of PAR and present two variations of the Pedestrian Attribute (PETA) [Den14] and Richly Annotated Pedestrian v1 (RAPv1) [Li16a] datasets, both following the zero-shot setting on pedestrian identity. Specifically, new training, validation, and test splits are introduced that prevent individuals from appearing in multiple splits. Furthermore, the authors implement a strong baseline for PAR, demonstrating that achieving state-of-the-art performance is possible with a simple, global approach. Consequently, the model proves appropriate for real-world application and, thus, is utilized as the baseline in this thesis. Further information is presented in Chapter 5. The MSCC [Zho21] method is more complex than the strong baseline by Jia et al. [Jia21b], as it uses a multi-scale model with two sub-modules to enhance performance. The spatial calibrated module captures features across various receptive fields, facilitating the inclusion of contextual information for the attributes and enabling the interaction of context information across spatial scales. On the other hand, multi-scale feature fusion based on non-local blocks [Wan18b] is employed. Final attribute predictions are derived from the predictions at each scale and the global predictions by maximum pooling.

Part-based methods: Part-based models simultaneously incorporate local and global information by dividing the input into multiple regions. Manual partitions or external components are necessary to locate relevant part regions. Fine-grained local information may improve recognition performance,

especially for highly localized attributes such as glasses. However, it is crucial to consider several drawbacks. Accurately localizing parts is necessary but may be an additional source of error. Furthermore, the increased complexity leads to longer training and inference times. Part-based methods are distinguished based on the approach used to locate the parts. The straightforward approach of dividing an input image is to manually define multiple regions of interest. Zhu et al. [Zhu15] leverage a predetermined grid to partition the image into 15 overlapping patches. Deep features are extracted from each local part. These parts are assigned to particular attributes according to their relevance to classification determined in advance. For instance, global attributes such as gender are determined by gathering information from all patches, while local attributes such as hairstyle rely only on information from local parts in the uppermost region of the input image. However, misaligned bounding boxes may cause serious issues as the manually defined local patches may not properly display the corresponding attributes and may include background clutter. Automatic approaches that specifically localize relevant parts in the images are expected to be more precise. One direction followed in related research is the application of object detection methods to identify relevant body parts. Early research focuses on poselets [Bou11] to decompose the human body into multiple parts. For instance, the PANDA [Zha14a] model extracts separate feature representations for each poselet in addition to a global feature of the entire image using a CNN. The fused global and local representations serve as input for an Support Vector Machine (SVM) in order to perform the attribute classification. In contrast to PANDA, Gkioxari et al. [Gki15] employ a deep version of poselets that replaces the commonly used low-level gradient orientation features with deep features extracted through a CNN. Analogous to PANDA, the approach then generates part-level features for the head, torso, and legs, and a holistic feature for the entire person. Classification is performed using a linear SVM. Li et al. [Li16b] omit conventional methods such as SVM and instead rely on a fully neural network-based approach. Fast R-CNN [Gir15] is employed to detect the entire persons as well as five specific body parts within full frame

images. Two additional branches are introduced, which consider the human-centric and scene context, in addition to the model branches that predict semantic attributes for the body parts and the overall appearance of the person. The human-centric context takes into account the nearest neighbors of body parts within the scene and the scene context the entire frame to rectify visual ambiguities. Furthermore, the scene context is utilized to classify actions. The LGNet [Liu18b] applies a proposal generator instead of detecting a predefined set of local regions. This produces a large number of proposals from which local features are extracted. These local features are then weighted separately for each attribute based on whether the proposal areas include high activation regions from a global attribute recognition branch. An alternative to the use of object detection to localize relevant local parts is the utilization of pose estimation methods. Yang et al. [Yan16] propose an end-to-end framework that is capable of learning PAR and part localization jointly. For part localization, the model first locates the key points of the human body to derive head, torso, and leg regions. The resulting part bounding boxes are refined by estimating adjustment parameters. Classification is performed by fusing the part features and feeding the resulting representation into a Fully-Connected (FC) network. The PGDM [Li18] approach addresses the problem that pose annotations are rarely available for surveillance datasets. Therefore, the model aims to distill the information from a pre-trained pose estimator and adaptively localize relevant parts of the image with only image-level supervision. In this case, separate attribute predictions are produced for the local parts and the entire person images and finally fused. In contrast to the aforementioned approaches, DeepCAMP [Dib16] uses pattern mining to identify and refine relevant image parts based on mid-level features. After extracting initial discriminative patches for each attribute, they are iteratively optimized by updating the patch clusters and retraining the classification network.

Attention: In addition to coarse human body parts, research explores attention modules for automatically discovering discriminative regions within images. The HydraPlus-Net [Liu17] exploits attention at various semantic feature levels in addition to a global feature branch. The authors' main contribution is the use of multi-directional attention, which applies attention not

only at the current feature stage but also at adjacent feature levels. In a similar approach, Sarafianos et al. [Sar18b] utilize attribute-agnostic attention at multiple scales. Attribute predictions are separately calculated for each feature level and the global branch. While the HydraPlus-Net combines the features and forwards them through a FC classification layer, Sarafianos et al. [Sar18b] fuse the predictions. The ALM [Tan19c] method also utilizes attention at multiple stages of the model, but in contrast to the previous methods, attribute-specific attention is used to locate relevant regions and make separate predictions at each scale. Analogous to Sarafianos et al. [Sar18b], the predictions from each scale are combined with global predictions to obtain the final attribute predictions. Guo et al. [Guo17b] employ an attention heat map refinement module with an additional loss function to refine the model's attention. The weights of the FC classification layer of the global attribute prediction branch are used to linearly combine the last feature maps generated by the backbone model, creating attention heat maps. An exponential loss function is implemented to concentrate the attention heat maps on smaller, more focused regions, as the authors claim that such heat maps are more appropriate for recognizing semantic attributes. Guo et al. [Guo19] discovered that attention maps generated by CNNs are inconsistent when the input image undergoes spatial transformations, such as horizontal flipping. This contradicts human perception and lacks intuition. As a solution, the authors introduce the VAC method, which enforces consistency in attribute predictions and attention heat maps across different spatially transformed variants of the input image by applying the Mean Squared Error (MSE) loss. The SSC approach, as proposed by Jia et al. [Jia21a], also addresses inconsistent attention heat maps, but in this case across different images. The authors argue that regions associated with attributes generally appear in similar parts of images. A spatial memory aggregates reliable class activation maps across images and serves as ground truth to enforce the spatial consistency of attention maps for the attributes. Similarly, semantic consistency of features is ensured by extracting attribute-related features using the class activation maps and applying a regularization loss to the resulting features with supervision from a semantic memory. Contrary to the previous approaches, the VeSPA [Sar17] model learns attention regarding the viewpoint of the person rather than spatially

localizing attributes. This is motivated by the fact that attributes, such as hair or backpack, appear differently from the front, side or back. To address this, the model incorporates separate classification heads for different viewpoints. The final attribute predictions are calculated by weighting and combining the resulting predictions of the heads based on the outputs of a view prediction module. Tan et al. [Tan19b] employ three separate branches, each incorporating a unique attention mechanism. Label attention is responsible for localizing discriminative, attribute-specific features. Spatial attentions attempts to globally identify relevant image regions for all attributes. Last, parsing attention leverages a human parsing network to achieve pixel-wise assignment of human body regions. The attention module divides the features of different body regions and subsequently combines them with convolutions in order to learn the location of discriminative parts and the optimal aggregation of the features simultaneously.

Loss functions: Further research investigates the influence of adjustments to the loss function on the PAR task. Zhou et al. [Zho17] propose a new weighted cross-entropy loss function to address the problem of imbalanced attribute distributions, based on the positive ratio of attributes in the training set, similar to the approach of Li et al. [Li15]. Moreover, Yang et al. [Yan20] introduce a hierarchical feature embedding approach that utilizes attribute as well as person identity information. The HFE loss enforces proximity of features for images showing the same attribute as well as for images depicting the same individual. This improves the embedding of challenging samples, such as those with barely visible backpacks, thereby enhancing the robustness against adverse conditions. In addition to the MTA-Net, a sequential attention approach to PAR, Ji et al. [Ji20] suggest combining the focal loss function with a weighting mechanism to address imbalanced attribute distributions. This method increases the focus not only on rare attributes but also on attributes that are difficult to recognize despite sufficient training samples. Zheng et al. [Zhe21] also utilize the focal loss function alongside a multi-label contrastive loss to obtain discriminative features. The contrastive loss aims to minimize the gap in predictions between samples that share the same attribute or lack a particular attribute, while simultaneously enlarging the difference for attributes with opposing labels among the samples.

Transformer-based methods: A recently emerging field of research is the use of transformers for PAR. The VTB [Che22b] method utilizes transformers to encode the set of predictable attributes into textual features before training. This approach captures the semantic correlations between attributes. Moreover, a cross-modal feature fusion technique combines these textual features with visual features extracted from the input image by employing a transformer encoder. The resulting visual-textual features serve as input to independent classification networks for each attribute. The authors claim that the VTB model more effectively captures both intra- and cross-modal correlations compared to alternative methods. The DRFormer [Tan22] adopts the same idea in a pure transformer-based approach. Initially, a transformer encoder is utilized to process the input image and generate global attribute predictions and spatial token features. Attribute-specific spatial embeddings are generated by applying attention to the token features. Subsequently, these embeddings are fused with semantic embeddings, acquired through BERT [Dev19] for the set of attributes, in a second transformer encoder to capture spatial and semantic relations. Fan et al. [Fan23] propose the PARFormer, which does not process semantic embeddings and is designed as a transformer-based baseline for PAR. The use of transformer is motivated by the fact that multi-head self-attention modules can capture long-range dependencies from a global perspective. This ability is restricted in CNNs due to the limited receptive fields of the convolutions. Besides, the authors propose a data augmentation strategy to improve the learning of attentive features, adapt the center loss for the multi-attribute task, and introduce an additional loss function to leverage viewpoint information.

Video-based PAR: In contrast to image-based PAR, video-based PAR aims to generate track-level attribute predictions. Video-based approaches process multiple images of a person over time, enabling rich capturing of visual appearance information compared to individual images. The lack of appropriate datasets has prompted research to extend the Motion Analysis and Re-identification Set (MARS) [Zhe16] dataset with semantic attribute annotations, as done by Chen et al. [Che19]. This dataset is utilized for the video-based experiments in this thesis. Details about the dataset are provided in Section 4.1.6. In addition to the annotations, Chen et al. [Che19] present a

convolutional temporal attention method for video-based PAR. The backbone model initially generates frame-level features. Subsequently, the model is divided into two parts: one part recognizes the person-dependent semantic attributes and the other classifies identity-irrelevant information regarding the pose and motion of the depicted individual. Temporal attention is learned separately for each attribute as well as for pose and motion to determine the significance of frames. Zhu et al. [Zhu23b] consider the task as a visual-textual feature fusion task and leverage pre-trained CLIP models to extract visual features from input frames and textual features for the set of relevant attributes. The model performs the final fusion of visual and textual features for attribute recognition using a transformer. Apart from these few task-specific methods, several types of general video-processing approaches are applicable. For instance, recurrent models, such as the CNN-RNN [McL16], can be employed. Separate features are extracted for each frame that function as input to a recurrent layer that captures the temporal context. The final track-level feature is acquired by temporally pooling the features from each time step. Another commonly used methodology in this field is the 3D neural network, which extracts spatial and temporal information in a single forward pass without any recurrent components. Various configurations have been explored in the literature, such as complete 3D models that solely contain 3D convolutional layers [Tra15, Har18], MCx [Tra18] models that comprise 3D convolutions in the initial stages and 2D convolutions in the latter stages, as well as (2+1)D [Tra18] networks that use 2D convolutions followed by temporal 1D convolutions to reduce computational complexity. A further approach to video processing are non-local blocks [Wan18b], which are convolutional components that learn filter offsets for across space and time. Thus, this approach captures long-range dependencies and contextual information effectively. Non-local blocks are initialized in the same manner as normal convolutions, allowing for seamless integration into existing CNNs. In recent times, transformer-based models have proven to be effective in processing sequential data. A representative and universally applicable model within this category is the popular VTN [Nei21] method. Initially, spatial features are extracted by a 2D model separately for each input frame, analogous to the CNN-RNN [McL16] and

temporal attention approach by Chen et al. [Che19]. Subsequently, a transformer encoder is implemented for temporal attention and feature fusion.

2.2.3 Learning a Joint Feature Space

Recent works [Jeo21] argue that attribute-based person retrieval using PAR methods is unreliable due to the challenging nature of PAR itself. Variations of the attributes' appearances or images with poor quality might impair recognition accuracy which results in imperfect retrieval rank lists. The alternative is to learn a cross-modal feature space between attribute categories, *i.e.*, different combinations of attributes, and images in which corresponding embeddings are close to each other while category-image pairs not belonging together are further apart. The AIHM [Don19] model learns joint hierarchical embeddings. Global category-level textual-visual embeddings as well as local embeddings on attribute-level are aligned and fused by a matching network that outputs the similarity score. Similarly, the TAVD [Iod20] framework also aligns global textual and visual representations as well as embeddings on the attribute level. For the latter, global visual features are decomposed into attribute-specific representations. Yin et al. [Yin18] and Cao et al. [Cao20] bridge the modality gap in the textual-visual feature space by adversarial optimization. In contrast to Yin et al. [Yin18], which consists of a single Generative Adversarial Network (GAN) for modality alignment, the SAL [Cao20] approach uses a second GAN to synthesize features of attribute combinations not included in the training set. The limited number of different sets of attributes in training data and, therefore, many unseen combinations is one of the primary issues with learning a joint textual-visual feature space. Enough variations are required to train the attribute encoder sufficiently to produce meaningful embeddings in the higher dimensional joint feature space, especially for those unseen combinations. Methods from the literature tackle the problem by mining new person categories and including them in the training set as negative examples [Don19], applying additional regularization techniques to enhance semantic consistency [Yin18, Jeo21], or using GANs for generating synthetic features [Cao20]. In contrast, PAR-based approaches treat the attributes independently and, thus, are able to correctly

recognize new attribute sets without the need of addressing this issue in particular. Adversarial training as applied in [Yin18, Cao20] tends to be unstable w.r.t. convergence due to the min-max optimization procedure. Thus, Jeong et al. [Jeo21] introduce novel loss functions with a conventional training scheme that outperform adversarial methods. The first loss function builds on the ArcFace [Den19] loss and pulls corresponding image and attribute embeddings closer together, while simultaneously increasing the distance between image embeddings and the embeddings of irrelevant attribute combinations. Additionally, the authors propose the ASMR regularization technique which assures that distances between embeddings follow their semantic relations in the binary attribute space. Therefore, the margin of semantically similar samples should be closer than those of samples with fewer attributes in common. Zhu et al. [Zhu23a] also propose a new loss function that considers both inter-modal as well as intra-modal matching. Specifically, embeddings are aligned across the textual and visual modalities, and within each modality using triplet losses with hard sample mining to obtain more robust cross-modal feature representations. Additionally, a regularization technique is employed to ensure consistent differences between features and matching behavior, regardless of the modal configuration.

2.2.4 Summary and Discussion

This thesis focuses on the use of discrete attribute queries for person retrieval instead of natural language requests, due to their better suitability for the surveillance task. Specifically, it is decided to rely on PAR methods to compare queries and gallery samples in the semantic domain. Within the PAR research field, researches explore several directions, including global image-based models, part-based models, attention methods, and transformers. Many architectures extend global models and leverage additional features [Sar17, Sar18b, Tan19b], *e.g.*, local or attentive features, to improve the accuracy. This implies that global representations are vital to achieve strong PAR and thereby attribute-based person retrieval performance. Additionally, global models are more efficient in computation, which is important for deploying models in

real-world scenarios. Therefore, this thesis aims at enhancing global image-based methods in order to maintain the benefit of fast inference. Furthermore, the literature indicates that these models are easily extendable if necessary. Moreover, publications from the literature support the hypothesis that simple global models are capable of achieving state-of-the-art performance with results comparable to more complex approaches [Gki15, Jia21b, Fan23].

Concerning video-based PAR, multiple suitable approaches are identified, including temporal attention [Che19], 3D CNNs [Tra15, Tra18], and transformers [Nei21]. These methods effectively capture the temporal context. However, the argument presented here is that the temporal context is less significant in PAR, since semantic attributes typically remain static and their recognition is usually independent of movements.

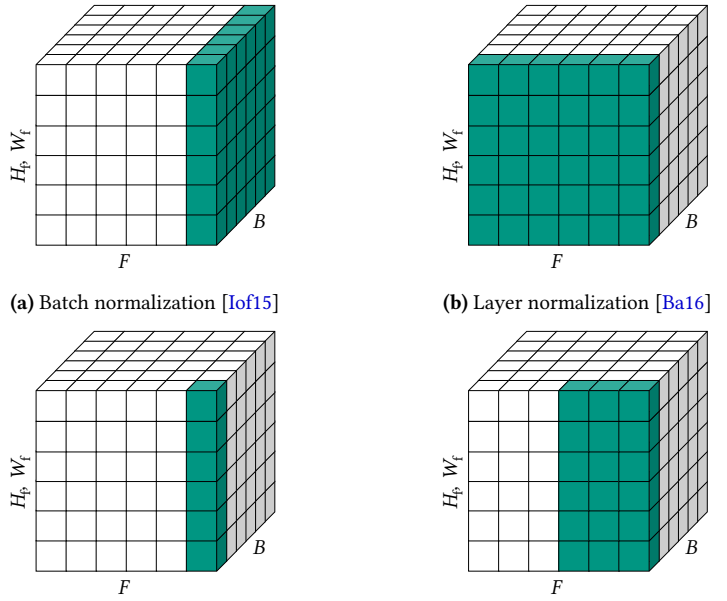
2.3 Further Related Literature

This section introduces further research from the literature that is not directly connected to attribute-based person retrieval but is relevant to certain optimizations proposed in this thesis. Concretely, a normalization module for PAR (see Section 6.2) is developed as well as a HP approach (see Section 7.1) that aims at determining the difficulty of recognizing semantic attributes in an input image. Background on literature concerning these topics is presented in Sections 2.3.1 and 2.3.2, respectively.

2.3.1 Normalization Techniques

Multiple normalization techniques have been developed in the field of deep learning. The primary purpose of each method is to standardize the considered features to have a mean of 0 and a variance of 1. Typically, additional shift and scale parameters are learned to enable adaptation to specific feature distributions and tasks. The approaches differ in terms of the dimensions taken into account, as illustrated in Figure 2.4. While F denotes the number

of feature channels, H_f and W_f stand for the height and width of a feature map, respectively. B signifies the number of images within a batch.



(a) Batch normalization [Iof15]

(b) Layer normalization [Ba16]

(c) Instance normalization [Uly16]

(d) Group normalization [Wu18]

Figure 2.4: Comparison of normalization layers – The figures shows different normalization layers that are frequently applied in deep learning models. F, H_f, and W_f represent the number of feature channels, feature map height, and width, respectively. B stands for the batch size. Changed representation after [Qia19].

Batch normalization [Iof15] standardizes the features separately per channel across the batch dimension to restrict changes in feature distributions during training, thus accelerating the training process. However, the major drawback is its dependence on the batches. To compute representative batch statistics during training, sufficiently large batches are required. Furthermore, during deployment, the learned statistics remain fixed, *i.e.*, batch normalization operates differently during training and testing phases. Layer

normalization [Ba16] eliminates the reliance on batches by normalizing activations along the feature dimension. In other words, the features of single training samples are normalized, allowing identical computation during training and test time. Instance normalization [Uly16], proposed as replacement for batch normalization in image stylization tasks, focuses on individual instances rather than entire batches or features. As a result, it removes contrast information from the feature maps. Last, group normalization [Wu18] resembles layer normalization, as it is employed along the feature dimension. Multiple groups of features are formed and normalized separately. The number of groups represents a hyperparameter of this technique.

In the context of PAR, Zhao et al. [Zha18b] suggest applying batch normalization to the model's output logits for enhanced recognition of imbalanced attributes in the training data. Batch normalization standardizes the output distributions for each attribute to avoid output distribution shifts that predominantly favor the most frequent attribute manifestation in the training data. However, normalizing solely in an attribute-wise manner may prove inadequate in the context of attribute-based person retrieval since it focuses on individual attributes, disregarding the entire person's attribute description. Thus, an improved normalization module for PAR and attribute-based person retrieval is developed in Section 6.2.

2.3.2 Hardness and Failure Prediction

Hardness or failure prediction is a rarely studied topic in the literature. Nevertheless, there are some similar works in motivation and methodology to the HP approach proposed in Section 7.1.

Zhang et al. [Zha14b] utilize a SVM with multiple kernels for 14 different conventional image features to predict failure independently of the classifier. Similarly, Daftry et al. [Daf16] also employ an SVM for failure prediction but apply it to deep features extracted by CNNs. The input image and optical flow

are encoded to base the failure prediction on spatio-temporal feature representations. Ramanagopal et al. [Ram18] also leverage spatial and temporal information to identify potential failure of object detectors in autonomous driving. Inconsistencies in detection results across stereo images and over time are evaluated. Another approach, proposed by Wang et al. [Wan17], involves a cascade of classifiers that are sorted by their accuracy and computational complexity. When a classifier is uncertain about an image's classification, the sample is passed to the subsequent stage, which contains a more precise but more complex classifier. Heavy models are only applied to the most challenging examples, thereby speeding up the classification process of easy samples. Further research adheres to the principle of learning to defer [Mad18, Moz20], incorporating deferral to more complex models through a cost term within the learning process. Sun et al. [Sun19] propose a lightweight method for detecting when a computer vision system is applied outside of its specified conditions, meaning conditions that deviate from the data distributions it was trained on. The HP approach presented in this thesis is based on the realistic predictor concept introduced by Wang et al. [Wan18a]. The realistic predictor includes a separate CNN utilized for difficulty prediction, thereby identifying classification failure. To ensure a certain level of classification accuracy, the authors suggest disregarding samples with hardness estimates above a designated threshold. Recently, researchers have expanded the concept of realistic predictors to further research fields such as regression and semantic segmentation [Gad23]. The FSNet introduced by Rahman et al. [Rah22] leverages mid-level features in addition to the input image to predict the difficulty of semantic segmentation at the pixel level. Contrary to previous methods, some publications do not utilize separate difficulty prediction models or branches. Instead, the methods rely on the variance of gradients [Aga22] or the investigation of loss values [Arr23].

Besides, additional research areas may be adopted to the objective of hardness prediction, encompassing uncertainty estimation methods [Gal16, Mad19, Gaw21], model calibration techniques [Guo17a, Wan23a], and approaches related to open set classification [Gen21, Mah21].

In this thesis, the widely-used approach of applying a separate module for difficulty prediction is pursued, since the related literature shows promising results and applicability to a variety of applications.

3 Concept

This thesis focuses on the development of a deep learning-based framework specifically designed for attribute-based person retrieval in real-world video footage captured by large-scale multi-camera networks. The objective is to create a robust system that can be seamlessly integrated into smart city or surveillance applications. The methodology proposed in this thesis serves as a vital search component, enabling efficient retrieval of individuals based on a description of their semantic attributes.

In general, several approaches are suitable for discrete attribute-based person retrieval, as introduced in Chapter 2. In this work, the concept of leveraging PAR to extract the semantic attributes of individuals included in the gallery and determine the similarity to the query in the attribute space is followed due to multiple reasons:

- **Explainability and interpretability:** PAR methods directly extract semantic attributes of depicted individuals, providing information in a human understandable and interpretable manner.
- **Flexibility:** Extracting semantic attributes offers flexibility in making requests for attribute subsets. Querying specific combinations of attributes is straightforward, allowing for tailored and customized retrieval requests.
- **Deployment:** Soft biometric attributes are independent of the deep learning model. Thus, model updates are smoothly applicable without the need to re-compute existing feature databases.
- **Training data requirements:** PAR methods have less requirements concerning training data compared to learning a joint visual-textual

feature space, as the lack of diverse attribute combinations in existing datasets has a minor impact.

- **Faster search:** PAR-based approaches eliminate the need for extracting query features during the search, as the matching occurs in the semantic attribute space. Additionally, pre-computed attribute vectors generally have lower dimensions compared to abstract appearance feature vectors usually learned in a common visual-textual feature space [Cao20, Jeo21, Zhu23a], accelerating distance computations.
- **Complementary information:** Soft biometric attributes are complementary information to related tasks such as person re-identification and may improve the performance of such tasks [Lin19, Spe20b].

Figure 3.1 illustrates the complete processing framework involved in the system. It outlines the various stages from camera streams or videos to the generation of search results to offer insights into the context of the methods explored throughout this thesis.

First, movements of persons within the entire multi-camera network are tracked, which serves multiple purposes. Aggregating the occurrences of the same individual reduces the size of the gallery database and simplifies the resulting rankings, since tracking avoids receiving a multitude of search results for the same individual at different times and locations. Furthermore, access to context information about the movements of persons or observed interactions is directly provided and richer information through multiple view of a person may increase the robustness of PAR. Afterward, a PAR model is applied to extract the soft biometric characteristics of the depicted individuals. Alongside the recognized attributes, hardness scores extracted by a HP branch are stored in the gallery database, providing complementary information about the reliability of the outputs generated by the PAR model. Within the framework, the gallery is queried using discrete attribute requests. Distances between queries represented as binary vectors and database entries are calculated and sorted to generate similarity rankings. Additionally, the

concept takes into account the evaluation part to assess the framework's suitability for real-world application.

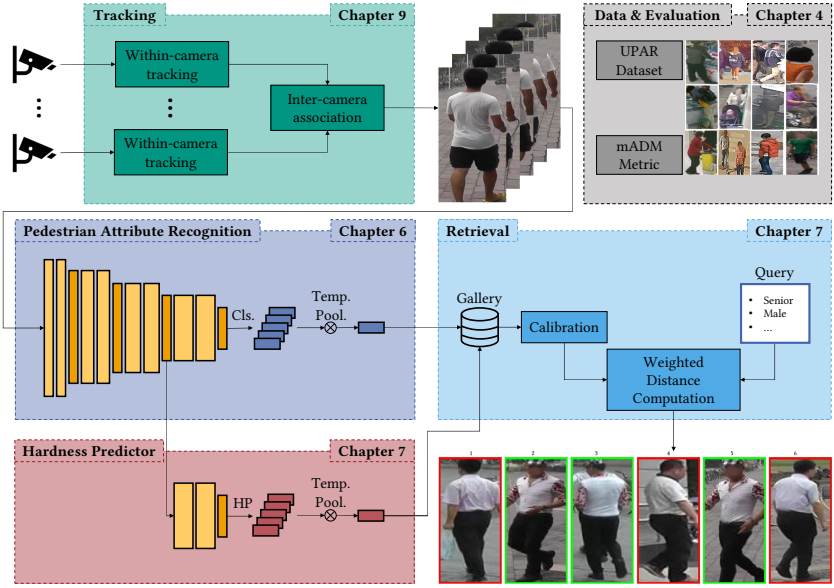


Figure 3.1: Concept overview – This thesis explores a holistic framework for attribute-based person retrieval. The tracking component processes input videos to aggregate data for individuals. Subsequently, semantic attributes of the persons are extracted using PAR methods, serving as the basis for retrieval. Additionally, a separate HP branch provides complementary information about the difficulty of classifying the attributes. Extracted data are stored in a database that can be queried using person descriptions to obtain a sorted ranking of gallery samples. Furthermore, a novel dataset and evaluation metric are introduced to facilitate meaningful evaluation of the task.

The primary focus of this thesis is on the PAR and retrieval stages, as they represent the core components of the framework. Nevertheless, research in terms of MTMCT is briefly discussed in Chapter 9 to provide necessary context for the application in real-world scenarios. The remainder of this chapter elaborates on the fundamental concepts and principles behind the contributions made to each specific stage in the order of appearance in this thesis.

Evaluation: To develop and evaluate data-driven methods for a specific use case, it is crucial to identify meaningful data and evaluation metrics for the task. If aiming at applying models in a real-world scenario it is important to learn features that are not overfitted to specific sceneries, person appearances, or dataset biases. Measuring the generalization performance is only possible if training and test domains originate from different sources. Nevertheless, none of the existing research datasets offer such a scenario [Den14, Zhe15, Li16a, Liu17, Li19a]. Most of the datasets lack diversity regarding, *e.g.*, demography, lighting conditions, and variance in clothing, which leads to models with poor generalization abilities. Some datasets even contain images of the same individuals in both the training and test sets [Jia21b]. Cross-dataset evaluation is impossible due to minimal intersections of attribute annotations in the datasets. To close this research gap, the UPAR dataset [Cor23, Spe23b] is proposed that harmonizes existing datasets. Thus, this allows for the first time to investigate the generalization performance of PAR and attribute-based person retrieval methods. Besides, typically used evaluation metrics for retrieval tasks in general only consider search results as relevant or irrelevant. However, this procedure ignores that search results can match a query to different extents in the attribute-based person retrieval case. Specifically, it disregards the ratio of matching attributes between the query and a gallery sample. To address this limitation, the Mean Average Degree of Match (mADM) measure is proposed as an additional evaluation metric to provide an evaluation score that is consistent with the visual quality of retrieval rankings as perceived by a system operator [Spe23a].

Pedestrian attribute recognition: The PAR model serves as the feature extractor in the proposed system, responsible for recognizing soft biometric features that are compared to discrete attribute queries in the retrieval stage. Thus, its role in achieving good person retrieval performance is crucial. This thesis relies on a global image-based approach due to the favorable tradeoff between accuracy and inference time. Additionally, research from the related literature suggests that it is possible to achieve state-of-the-art results with such models despite the lightweight architecture [Gki15, Jia21b]. While previous works have mainly focused on the PAR task itself, the importance of

its usefulness for downstream tasks such as retrieval has been mostly overlooked. Therefore, this thesis systematically examines and evaluates design choices for creating a robust attribute classifier, considering their impact on attribute-based person retrieval for the first time [Cor23, Spe23b]. Furthermore, this investigation aims to identify weaknesses and to propose solutions to reduce their influence. One such weakness are the imbalanced training distributions of attributes, which lead to biased predictions favoring the most common attribute manifestation during deployment. The most common measure against this issue is to increase the focus on training samples with rarely occurring attributes. However, this may result in an overestimation of recognized attributes per instance and an overconfident model, which likewise deteriorates the retrieval results. To mitigate these issues and to find the optimal tradeoff, a proposed normalization module calibrates the attributes logits both attribute- and instance-wise, simultaneously improving PAR and retrieval performance [Spe23a]. In real-world applications, video-based recognition of soft biometrics is required. While there are various approaches for video data processing in the literature, this work argues that a temporal pooling approach is sufficient for PAR since relevant attributes are not dependent on movements [Spe20c]. In addition, this approach offers greater flexibility and lower computational complexity compared to techniques such as 3D CNNs. Detailed comparisons are conducted to validate the hypothesis.

Hardness prediction: Despite the advancements in attribute recognition described above, non-determinable attributes, for instance, due to occlusions of body parts or misaligned bounding boxes pose a significant challenge for attribute-based person retrieval. Figure 3.2 visualizes exemplar images for which not all attributes are decidable. These attributes are still classified by the PAR model and utilized during retrieval, even though they cannot be accurately determined. Consequently, the lack of crucial information leads to errors in the retrieval process, where individuals matching the query description are mistakenly considered irrelevant. To tackle this issue, an independent HP is proposed. This predictor provides a separate difficulty score for each attribute, indicating the likelihood of uninformed decision of the classifier [Spe20a]. By incorporating these hardness scores as weights during the retrieval, attributes with high certainty are prioritized over undecidable

attributes [Flo21]. This effectively minimizes the negative impact of invisible and indeterminable attributes on the retrieval performance, ensuring the real-world search results are more accurate and reliable.

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

Figure 3.2: Examples for indeterminable attributes – Certain attributes in the images are indeterminable due to view points (Figure 3.2a and Figure 3.2d), misaligned bounding boxes (Figure 3.2b), occlusions (Figure 3.2c), and low illumination and image resolution (Figure 3.2e). Image source: [Zhe15].

Person retrieval: To build the search ranking, the distance between recognized attributes and query attributes is determined. Typically, the Euclidean distance is calculated between binary query attributes and attribute confidences obtained through the PAR model for the gallery samples stored in the database. However, opportunities for enhancement exist due to often poorly calibrated PAR models, unbalanced output distributions of the classifier, and variations in recognition accuracies for the attributes. For instance, if a soft biometric characteristic is seldom present in the training set, the classifier is highly certain with low deviation if such an attribute is not present but produces scattered outputs for samples showing the attribute. The domain gap between training and test data aggravates the issue. These factors cause deteriorated search performance, but are not addressed in related research concerning person retrieval with PAR methods. In this thesis, first, methods for reliability calibration [Pla99, Zad01, Luc18] of the PAR model's outputs are investigated to align the attribute confidences with the empirical probabilities for the attribute presence. Furthermore, a classifier is typically more uncertain for hard-to-recognize attributes and, thus, produces larger distances for such

attributes. Since this focus on challenging attributes during distance computation is unintended, the Euclidean distance computation is weighted by the average recognition errors of attributes to achieve equal contribution [Spe21a]. Last, it is proposed to directly leverage the output distributions of the classifier to compute a Distribution-Based Distance (DBD) in addition to the Euclidean distance [Spe21a]. This helps to further compensate for still remaining shifts of confidence scores for present and absent attributes.

Tracking: Finally, the integration of attribute-based person retrieval into an overall system for real-world applications is briefly presented. As there are no suitable large multi-camera person tracking datasets due to difficulties in creation, such as annotation effort and privacy regulations, the synthetic Multi-camera Track Auto (MTA) dataset is generated offering a diverse scenery and highly accurate annotations [Köh20]. Furthermore, methods for tracking people within a network of cameras are briefly introduced, focusing on the support of both use cases presented in Section 1.1: offline [Köh20] and online [Spe22a] investigations. The two MTMCT approaches utilize scene information to enhance the tracking accuracy. In addition, an exemplary implementation of the proposed framework is outlined to bridge the gap to real-world application and demonstrate its functionality and effectiveness [Spe22c].

4 Experimental Setup

In this chapter, the experimental setup is presented, which is used to evaluate the proposed methods in the context of this thesis for PAR and attribute-based person retrieval. First, an overview of the utilized PAR datasets is provided in Section 4.1. Then, evaluation metrics are discussed in Section 4.2, followed by the introduction of the evaluation protocols in Section 4.3.

4.1 Datasets

An overview of publicly available datasets in the surveillance domain with soft biometric annotations is offered in Table 4.1. In general, these datasets are categorized into two types: image and video datasets. Image-based datasets provide single images of individuals, whereas video-based datasets supply either full-frame videos with track annotations or person tracks containing cropped bounding boxes from subsequent frames for each person.

A few image datasets depict large-scale scenarios, however, several datasets do not contain sufficient person images to be suitable for evaluating the methods discussed in this thesis. For instance, APiS [Zhu13], VIPeR [Gra07, Lay12, Lay14], GRID [Loy10, Lay14], and SoBiR [Mar16] datasets have fewer than 4,000 images, which is hardly adequate for training and evaluating deep learning models. Similarly, further datasets such as PRID [Hir11, Lay14] lack diversity in terms of viewing angles and scenery since the dataset was collected using only two different cameras. Based on these consideration, Market-1501 [Zhe15, Lin19], PETA [Den14], PA-100K [Liu17], and RAPv2 [Li19a] datasets are chosen as benchmarks for the experiments in this thesis. Each of these datasets is introduced in detail in the following sections. Additionally,

the UPAR dataset is proposed to harmonize these existing datasets, increasing diversity and enabling comprehensive generalization experiments concerning PAR and attribute-based person retrieval.

Table 4.1: Overview of PAR datasets – The table shows publicly available research datasets for PAR, which are also suitable for attribute-based person retrieval. Only a few datasets offer sufficient diversity and size to train modern deep learning models and to reflect realistic scenarios. †The datasets consist of the entire videos or frames. Resolution refers to the resolution of the video frames.

Type	Dataset	Number of Persons	Number of Images	Resolution	Number of Attributes	Number of Cameras
Image	APiS [Zhu13]	–	3,661	48×128	35	–
	VIPeR [Gra07, Lay12, Lay14]	632	1,264	48×128	21	2
	PRID [Hir11, Lay14]	934	24,541	64×128	21	2
	GRID [Loy10, Lay14]	1,025	1,275	29×67 to 169×365	21	8
	PETA [Den14]	8,705	19,000	17×39 to 169×365	105	–
	Market-1501 [Zhe15, Lin19]	1,501	32,217	64×128	30	6
	SoBiR [Mar16]	100	1,600	60×150 to 191×297	12	8
	PA-100K [Liu17]	–	100,000	50×100 to 758×454	26	598
	RAPv2 [Li19a]	2,589	84,928	33×81 to 415×583	72	25
Video	MARS [Zhe16, Che19]	1,251	1,191,003	128×256	52	6
	SAIVT† [Hal18]	151	–	704×576	16	6
	P-DESTRE† [Kum21]	269	–	3840×2160	16	–

Furthermore, this work studies approaches for video-based PAR, as video processing is a requirement in real-world systems. For conducting the experiments, the MARS [Zhe16, Che19] dataset is chosen due to the inclusion of an substantial higher number of individuals. In contrast to the SAIVT [Hal18] and P-DESTRE [Kum21] datasets, MARS consists of cropped person bounding boxes and does not contain the entire videos or frames.

4.1.1 PETA

In their work, Deng et al. [Den14] address the limitations of small-scale datasets by introducing the first large-scale PAR dataset called PETA. This

dataset is a composite of ten smaller datasets, therefore, offering diversity in terms of scenarios, persons, and camera models. Example images from the PETA dataset are illustrated in Figure 4.1.

Figure 4.1: Example images from the PETA dataset – The PETA contains images from ten sub-datasets. Thus, the images depict diverse lighting conditions, indoor and outdoor scenes, and closely as well as widely aligned person crops.

The PETA dataset includes both indoor as well as outdoor imagery captured by static surveillance cameras. In total, it contains 19,000 images with annotations for 61 binary and four multi-class attribute annotations, which equals 105 binary soft biometrics. The dataset is partitioned into 9,500 images for training, 1,900 for validation, and 7,600 for testing, respectively. The resolution of the person crops ranges from 17×39 to 169×365 pixels. It encompasses a vast number of different individuals, totaling 8,705 identities.

However, a drawback of the dataset is that attributes were not annotated in an image-wise manner. Instead, attribute annotations were determined based on a randomly selected image for each individual. Consequently, the annotations might differ from the actual visual perception. Some attributes might not be visible but are annotated, while others that are visible might be ignored. Furthermore, the official evaluation protocol by the authors only utilizes 35 selected attributes for the experiments and disregards soft biometric features like *sunglasses*, which have few occurrences. The evaluated soft biometric characteristics include 15 of the essential attributes in video surveillance according to human experts [Lay12], along with 20 additional attributes that aroused interest of the authors. In this work, the same subset of 35 attributes is leveraged.

4.1.2 Market-1501

The Market-1501 [Zhe15] dataset was initially introduced as a person re-identification dataset and was captured on a university campus in China. It includes 1,501 different individuals, with a total of 32,668 person crops. Each of the images was resized to a standardized size of 64×128 pixels.

The individuals included in the dataset are divided almost evenly, with 751 identities assigned to the training set and 750 to the test set. This results in 12,936 training and 19,732 test images, respectively.

Soft biometric annotations for the Market-1501 dataset were contributed by Lin et al. [Lin19]. The annotations include 26 binary attributes and one multi-class attribute. The multi-class attribute, *age*, classifies individuals into four categories. Similar to PETA, soft biometric characteristics are annotated at the identity level with the identical limitations.

Furthermore, it is important to note that the dataset primarily represents a single limited scenario: summer on a Chinese campus. Consequently, an inherent bias toward young Asian people wearing summer clothing is present, as illustrated in Figure 4.2 showing sample images from the dataset.

Figure 4.2: Example images from the Market-1501 dataset – The dataset was collected near to a supermarket on a Chinese campus. Therefore, primarily young Asian people are displayed.

4.1.3 PA-100K

The most extensive PAR dataset to date, consisting of a total of 100,000 person images, is the Pedestrian Attribute 100K (PA-100K) dataset introduced by

Liu et al. [Liu17]. This dataset offers tremendous diversity concerning image resolution, lighting conditions, and environments since it was captured from 598 different outdoor surveillance cameras. For instance, the spatial resolutions of person images span a range of 50×100 to 758×454 pixels. The broad range of image resolutions, lighting conditions, and scenes is evident in the example images presented in Figure 4.3.

Figure 4.3: Example images from the PA-100K dataset – The PA-100K dataset is the largest PAR dataset to data. The samples illustrate the broad range of lighting conditions and image resolutions included in the dataset.

The dataset is structured with 80,000 images designated for training and 10,000 images for each validation and testing. Concerning semantic attributes, annotations cover 26 binary soft biometrics. However, the dataset lacks annotations for crucial characteristics such as the clothing colors.

4.1.4 RAPv2

Similar to the PA-100K dataset, Li et al. [Li19a] acquired the data for their Richly Annotated Pedestrian v2 (RAPv2) dataset using a real-world surveillance camera network, consisting of 25 different cameras. However, this dataset captures an indoor scenario within a shopping mall.

The dataset is an expanded version of the RAPv1 [Li16a] dataset. The resolution of the 84,928 person images spans from 33×81 to 415×583 pixels. Annotations are provided for a total of 72 attributes for each image, with 69 being binary and three multi-class attributes. Similar to the PETA dataset, the authors use a subset of semantic attributes for PAR, specifically evaluating 54 binary attributes. Due to the comparatively large number of annotated soft

biometrics and fine-grained distinction between, for instance, shoe and clothing types, a substantial amount of strongly imbalanced attributes is included. Hence, the dataset is particularly challenging. In addition to soft biometrics, the dataset includes annotations for environmental and contextual information, such as viewpoint, occlusions, and body part bounding boxes.

The dataset is divided into training, validation, and test sets, containing 50,957, 16,986, and 16,985 images, respectively. Regarding distinct person identities, 2,589 individuals are included.

The images included in the dataset, as visualized in Figure 4.4, predominantly exhibit well-illuminated scenes due to the artificial lighting. However, notable challenges arise from low resolution images, where highly localized attributes are hard to recognize.

Figure 4.4: Example images from the RAPv2 dataset – The dataset was recorded inside a shopping mall. As a result, most images are well illuminated. However, challenges include low image resolution and strongly imbalanced attributes.

4.1.5 UPAR

The lack of adequate datasets for PAR and attribute-based person retrieval presents a severe task-specific challenge due to several reasons, as highlighted in Section 1.2.2.

On the one hand, publicly available surveillance datasets represent only a small excerpt of the real world, resulting in substantial biases related to various characteristics, such as age, ethnicity, lighting conditions, and seasonal

clothing. Biases in the training data are captured by deep learning models during training and, thus, pose a serious concern as they could lead to discrimination against certain cultural groups in practical application. For instance, the Market-1501 dataset exhibits bias toward outdoor imagery of young Asian people during the summer. Furthermore, except for PETA, data collection was primarily conducted either indoors or outdoors, which limits the ability to train models that perform well under diverse lighting conditions.

As a consequence, experimental findings based on these datasets might not be universally applicable to various real-world scenarios but rather transfer to applications with similar conditions and characteristics. During training, the models might overfit the specific training data distributions, resulting in poor generalization to novel domains with divergent properties. This problem is exacerbated by the comparatively small number of training images, particularly in datasets such as Market-1501.

Besides, each dataset possesses its own unique set of soft biometric annotations, preventing multi-domain training and cross-domain generalization experiments. Utilizing multiple datasets in training proves beneficial in enhancing diversity and mitigating the biases mentioned earlier. Moreover, it remains a challenge to accurately assess the genuine generalization capabilities of methods due to divergent semantic attributes across datasets. Models cannot be trained on one dataset and evaluated on another with the exception of a few shared attributes such as gender. Consequently, quantifying the domain gap between different domains and evaluating the true impact of biases is elusive.

To close this research gap and enable research under realistic evaluation settings, the author of this thesis proposes the UPAR dataset. Two publications by the author are related to this dataset [Spe23b, Cor23]. The UPAR dataset shares its fundamental concept of combining multiple existing datasets with the PETA dataset but is substantially larger in scale and its primary focus lies in evaluating the generalization capabilities of deep learning methods concerning PAR and attribute-based person retrieval. The UPAR dataset combines various existing data sources. Multiple considerations were crucial for this decision. On the one hand, this procedure reduces the annotation effort

since some of the original annotations from the sub-datasets are reused. On the other hand, it is argued that the conjunction of existing dataset already provides sufficient diversity. Consequently, there is no need to collect additional data, complying with the principle of data minimization. Capturing realistic image and video data of people in a surveillance context poses enormous challenges regarding privacy, particularly concerning the collection of the consent of appearing persons. Several researchers did not collect consent in an appropriate manner, which lead to the withdrawal of the popular and widely utilized DukeMTMC [Ris16] and Celeb1M [Guo16] datasets. Synthetically creating data using video games such as GTA V[1] is widely used within the research community [Fab18, Köh20] but may lack diversity in terms of clothing and human models in the context of PAR. Concretely, the UPAR dataset is composed of the Market-1501 [Zhe15], PA-100K [Liu17], PETA [Den14], and RAPv2 [Li19a] datasets. Thus, it includes diverse image data spanning over a variety of scenes and persons with various sets of soft biometric characteristics. When considering the entire UPAR dataset, biases present in the individual sub-datasets become less pronounced and lose their significance. In total, the four sub-datasets contribute a sum of 224,737 images to the UPAR dataset.

The UPAR dataset offers annotations for 40 different semantic attributes. These attributes were selected based on relevant literature discussing the use of soft biometrics in crime investigations, which is reviewed in Section 2.1.2. Soft biometrics are categorized into four categories according to the taxonomy of Dantcheva et al. [Dan16] (see Figure 2.1): demographic, anthropometric, medical, and material and behavioral attributes. It was decided to exclude two classes in the annotations. Anthropometric and medical soft biometrics, including body geometry, facial geometry, and body weight, are of limited relevance for attribute-based person retrieval. These characteristics are rarely reported by witnesses and are subject to uncertainty due to the own-anchor effect and regression to the mean effect [Spo92, Fli86]. Therefore, the focus is on demographic and material soft biometrics. Related literature [Lin94] indicates that testimonies primarily contain general and clothing-related information. Consequently, demographic characteristics such as age and

[1] https://www.rockstargames.com/de/gta-v

gender are included as semantic attributes in the UPAR dataset. Age is categorized into three distinct manifestation, whereas gender is considered binary. It is noteworthy that the gender attribute refers to the perceived gender by a witness based on visual appearance, rather than biological sex. Moreover, various soft biometrics related to clothing are annotated. This includes lengths and colors of both upper- and lower-body garments, as well as the type of the lower-body clothing, categorized into *trousers and shorts* or *skirt and dress*. Clothing length is encoded binary as either short or long, and eleven different colors are distinguished for both upper- and lower-body clothes, along with an additional class for colors not included in the predefined list. Additionally, accessories such as different types of bags and the presence of headwear represent important material soft biometrics, and are, thus, included in the UPAR annotations. Local semantic attributes such as glasses and hair length are also considered, despite being subject to uncertainty when reported by witnesses. The objective is to increase the dataset's difficulty and enable comparisons of PAR methods concerning their ability to recognize small and imbalanced characteristics. Three hair lengths are specified: short, long, and bald, and annotations for two types of glasses, normal and sun, are given. Table 4.2 provides an overview of the attribute annotations contained in the UPAR dataset.

Table 4.2: UPAR attribute annotations – Overview of the attribute annotations provided for the UPAR dataset. Annotations include demographic and material soft biometrics.

Category	Age	Gender	Hair length	Upper-body clothing length	Upper-body clothing color	Lower-body clothing length	Lower-body clothing color	Lower-body clothing type	Backpack	Bag	Glasses	Hat
Attributes	Young Adult Elderly	Female	Short Long Bald	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Trousers&Shorts Skirt&Dress	Backpack	Bag	Normal Sun	Hat

The annotation process was conducted using the Antonn[1] tool with an additional validation phase to ensure that multiple annotators assessed the attributes for each image, improving consistency and minimizing annotation errors. Furthermore, the annotators were provided with clear instructions and sample images showing the attributes and edge cases. These were iteratively refined based on feedback during the labeling process. Additionally, *unknown* labels were implemented for the attributes to identify indeterminable or highly challenging images. These samples were evaluated separately following the annotation process. In this stage, any images that were found to be inappropriate for the task were excluded from the UPAR dataset. Otherwise, missing labels were assigned. In total, the UPAR includes 3.3 million novel binary attribute annotations.

In Figure 4.5, the number of samples per attribute within the UPAR dataset is presented. The dataset covers a wide spectrum of attribute distributions, covering both rare attributes, such as certain colors, and attributes present in the majority of the 224,737 images. As is typical, rarely occurring soft biometrics are prevalent in the dataset. However, even for the most imbalanced attribute, which is the lower-body clothing color purple, a notable count of 451 images are included. This substantially increases diversity concerning intra-class appearances, especially when compared to just 29 images showing this particular attribute in the Market-1501 dataset. Furthermore, biases within attribute distributions have been mitigated. For instance, the Market-1501 dataset displays short upper-body clothing in more than 94% of the images. In contrast, the positive ratio of this attribute is reduced to 42% through the combination with the further datasets in UPAR, aligning more accurately with the actual distribution of this soft biometric in real-world scenarios.

[1] https://www.iosb.fraunhofer.de/de/kompetenzen/bildauswertung/video-exploitation-systems/antonn.html

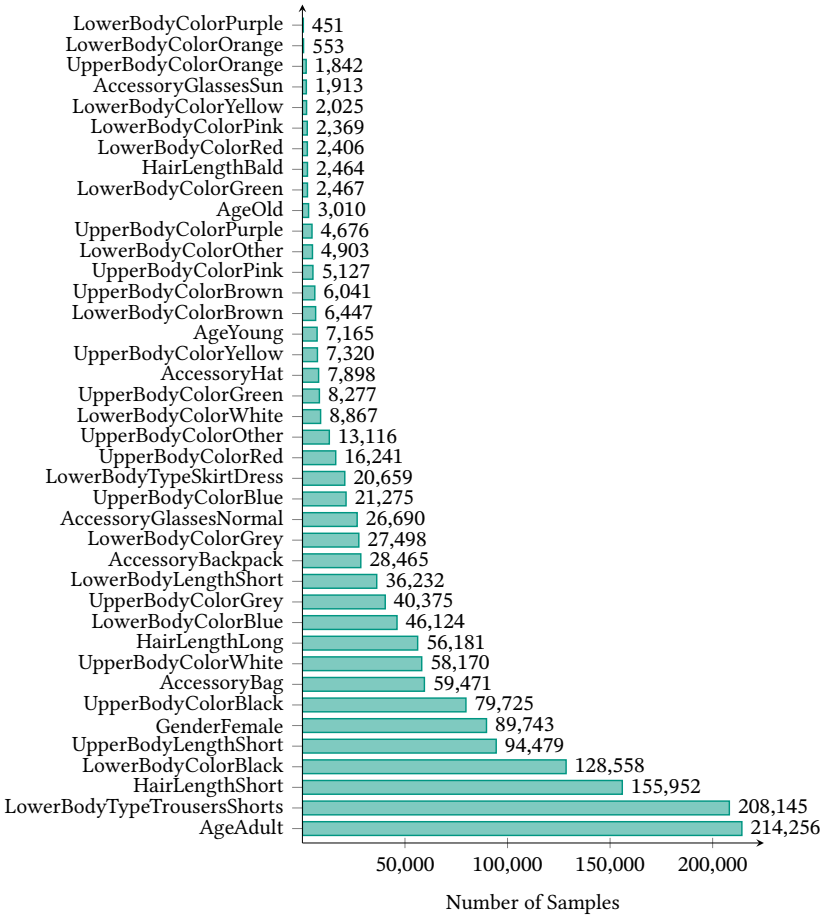


Figure 4.5: Number of samples per attribute in the UPAR dataset – The dataset covers a broad range of distributions of attributes. Notably, even the most imbalanced attribute with the fewest positive samples is depicted in 451 images.

In addition to providing annotations, the UPAR dataset proposes two evaluation protocols designed to assess the generalization capabilities of PAR and attribute-based person retrieval approaches. These protocols mimic realistic

deployment scenarios. The protocols are detailed in Section 4.3, where the evaluation procedures utilized in this thesis are introduced.

4.1.6 MARS

The MARS [Zhe16] dataset is a video-based dataset, which provides tracks of individuals, comprising a varying number of cropped bounding boxes.

Originally, the dataset was introduced for video-based person re-identification and is build upon the same data source as the Market-1501 dataset. Semantic attribute annotations were later contributed by Chen et al. [Che19].

Diverging from the Market-1501 dataset, the MARS dataset includes images of only 1,251 identities. These are split into 625 for training and validation, and 626 for evaluation. This corresponds to a total of 7,074 tracks for training and validation, and 6,848 tracks for testing, respectively. Each track comprises 15 to 920 cropped bounding boxes of individuals. On average, a track consists of 71.3 images. Similar to the Market-1501 dataset, images are scaled to an uniform size of 128×256 pixels.

In addition to annotations for common soft biometrics, such as clothing information, age, gender, and accessories, the annotations in the MARS dataset include motion and viewpoint details.

While the image-based datasets are employed throughout the entire thesis, the MARS dataset is specifically utilized in Section 6.3 to evaluate experiments concerning video-based PAR.

4.2 Evaluation Measures

In this section, the common measures typically used for evaluating the performance of PAR and person retrieval methods are presented. Additionally, the novel mADM [Spe23a] metric specifically tailored to the evaluation of attribute-based person retrieval is introduced. In this thesis, all evaluation results are presented as percentages, unless indicated otherwise.

4.2.1 Pedestrian Attribute Recognition

In general, two types of metrics are considered to evaluate PAR approaches: label- and instance-based criteria. Label-based metrics treat attributes independently. First, the scores are calculated for each attribute separately, and afterward, the average across the attributes is computed to obtain the final results. In contrast, instance-based, also referred to as example-based, metrics focus on inter-attribute correlations. Semantic attributes are not independent since some attributes greatly influence the prior probability of other attributes, *e.g.*, skirt and female or short lower-body clothing and short upper-body clothing. Thus, instance-based metrics are calculated concerning the attributes recognized in one image and averaged across the samples. Consistency of attributes recognized in a pedestrian image is measured, which is more meaningful for the attribute-based retrieval task. It is essential to get the best description of the person shown in the image instead of capturing single attributes independently. In this thesis, representatives of both types of metrics are applied.

Label-based Mean Accuracy (mA): The mA metric is a label-based evaluation criterion, which was originally adopted to the task of PAR by Deng et al. [Den14]. Contrary to the raw accuracy metric, mA considers the accuracy of positive and negative samples separately to deal with imbalanced attribute distributions. Otherwise, the trivial solution of always predicting the absence of attributes would lead to high accuracy scores since the majority of semantic attributes occurs rarely. Concretely, the mean value between the recall of positive and negative samples is calculated and subsequently averaged over the attributes as follows:

$$\mathrm{mA} = \frac{1}{2L} \sum_{j=1}^{L} \left(\frac{\mathrm{TP}_j}{\mathrm{P}_j} + \frac{\mathrm{TN}_j}{\mathrm{N}_j} \right). \tag{4.1}$$

L represents the number of attributes, while TP_j, P_j, TN_j, and N_j stand for the numbers of true positives, positive samples, true negatives, and negative samples for the j-th attribute, respectively.

Instance-based F1: The instance-based F1 score for PAR [Li16a] is based on the instance-based precision Prec_{PAR} and recall rate Rec_{PAR}. In contrast to label-based metrics, the metrics are computed separately per image $\mathbf{I}_i$ and then averaged across the M images. The calculation of the precision based on the positive ground truth labels and recognized attributes is formulated as

$$\text{Prec}_{\text{PAR}} = \frac{1}{M} \sum_{i=1}^{M} \frac{|Y_i \cap f(\mathbf{I}_i)|}{|f(\mathbf{I}_i)|}. \tag{4.2}$$

Y_i are the ground truth positive labels of the i-th example, $f(\mathbf{I}_i)$ returns the predicted positive labels for the i-th image, and $|\cdot|$ denotes the set cardinality.

Recall measures the fraction of attributes that are present in an image and are correctly recognized by the PAR approach. The definition of recall is provided in the following equation:

$$\text{Rec}_{\text{PAR}} = \frac{1}{M} \sum_{i=1}^{M} \frac{|Y_i \cap f(\mathbf{I}_i)|}{|Y_i|}. \tag{4.3}$$

The calculation is similar to precision, but the number of correctly recognized attributes is divided by the number of attributes that are present according to the ground truth annotations.

Finally, the instance-based F1 score F1_{PAR} is formulated as the harmonic mean between the precision and recall measures:

$$\text{F1}_{\text{PAR}} = \frac{2 \cdot \text{Prec}_{\text{PAR}} \cdot \text{Rec}_{\text{PAR}}}{\text{Prec}_{\text{PAR}} + \text{Rec}_{\text{PAR}}}. \tag{4.4}$$

4.2.2 Person Retrieval

To evaluate the performance of information retrieval systems, a test set is required that contains a set of so-called *documents*, a set of queries expressing the information needs, and ground truth relevance labels [Man09]. Relevance is usually considered binary, *i.e.*, *relevant* or *non-relevant*. In the case of attribute-based person retrieval, documents are equivalent to gallery images,

queries correspond to specific sets of binary attributes, and a person image is considered relevant if the depicted person's soft biometric traits exactly match the query description. In general, unranked and ranked retrieval results are distinguished. Since the objective in attribute-based person retrieval is to rank gallery images according to their distance to the query, only the ranked case is relevant for this thesis.

In this thesis, three evaluation metrics are employed for person retrieval, each with a different focus: Mean Average Precision (mAP), Cumulative Matching Characteristics (CMC), and mADM. The first two are well-established and widely used, while mADM is a novel metric introduced by the author of this thesis.

mAP: mAP is defined as the mean value of Average Precision (AP) scores across a set of queries. AP is an approximation of the area under the precision-recall curve and is defined as follows:

$$\mathrm{AP} = \frac{1}{\mathrm{GTP}} \sum_{k=1}^{|\mathcal{G}|} \mathrm{Prec}_{\mathrm{IR}}@k \cdot \mathrm{Rel}@k. \tag{4.5}$$

GTP refers to the number of ground truth positive samples in the gallery $\mathcal{G}$ for the respective query. $\mathrm{Prec}_{\mathrm{IR}}@k$ denotes the precision at a specific ranking position k, *i.e.*, the precision computed for the first k ranks. Together with the relevance indicator $\mathrm{Rel}@k$, which is 1 if the document at position k is relevant and 0 otherwise, AP is defined as the average of precision scores at ranking positions with relevant documents.

CMC: In contrast to mAP, CMC is no single figure metric. Instead, it focuses on the position of the first positive sample in the retrieval ranking, *i.e.*, it does not measure the quality of the entire ranking. The CMC top-k accuracy $\mathrm{Acc}@k$ denotes the frequency of positive matches in the first k ranks across all queries. The idea is that a retrieval system should provide positive matches in early ranks to improve its usefulness and reduce manual interventions and effort by the system operator. For a single query, the accuracy at rank k is

calculated as a shifted step function:

$$\text{Acc@}k = \begin{cases} 1, & \text{if a match is included in top-}k\text{ ranked gallery samples} \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

Finally, this accuracy score is averaged over the information needs, *i.e.*, the queries. This metric can be computed separately for each position in the ranking. In this thesis, it is focused on the accuracy at the first position of the rank lists. In the following, Acc@k is referred to as Rank-1 accuracy (R-1), since this is the commonly used term in the context of person retrieval.

However, there are some disadvantages to consider when using this metric. First, the measure is less stable than, for instance, mAP, since deviations concerning the first positive samples cause substantial differences in the resulting score. Furthermore, only the easiest match is considered if the gallery set contains multiple positive samples. Especially in attribute-based person retrieval that does not perform person identification and, thus, may have multiple identities matching a query, this has to be considered.

mADM: The aforementioned metrics assume binary relevance labels to compute the precision. This is appropriate in contexts such as person identification or re-identification, where it is unambiguously determinable if two images depict the same individual or not. However, in attribute-based person retrieval scenarios, attribute queries and attributes depicted in person images from the gallery may match to varying degrees. However, this fact is not addressed by current metrics. Considering the degree of similarity between gallery images and the query during evaluation is beneficial due to the following factors:

- Contrary to binary relevance labels, the degree of agreement provides complementary and more detailed information about the quality of the rank list.
- The degree of match is much more robust against annotation errors and negative influences by the use of instance-wise annotations, which are common problems when dealing with soft biometrics. While a wrong annotation concerning a single attribute leads to a

swap regarding the binary relevance, the degree of match is only slightly reduced. This also applies to cases where local attributes like glasses are invisible.

- Witness descriptions are prone to uncertainties. Therefore, individuals slightly deviating from the query description might nevertheless be relevant as search results.

To benefit from these advantages, a new evaluation metric is proposed that resembles mAP but incorporates the degree of agreement with the query. The measure is build on top of mAP and not CMC since person retrieval commonly addresses multiple identities and, thus, considering all matches in the rank list is important. For this, the $\text{Prec}_{\text{IR}}@k$ for a single query is modified to

$$\text{Prec}_{\text{DoM}}@k = \frac{1}{k}\sum_{l=1}^{k} \text{DoM}_{\text{Norm}}@l, \tag{4.7}$$

with $\text{DoM}_{\text{Norm}}@k$ being the normalized Degree of Match (DoM) between the query and the gallery sample at position k in the ranking. Given the number of attributes L and the Hamming distance d_k^{Ham} between the query and the ground truth annotations of the sample at rank k, the DoM, which is the basis for calculating the normalized DoM, for the single gallery sample at rank k is computed as

$$\text{DoM}@k = \frac{L - d_k^{\text{Ham}}}{L}. \tag{4.8}$$

$\text{DoM}@k$ quantifies the ratio of matching attributes between the query and a gallery sample. If all attributes between the query and the sample at position k agree, d_k^{Ham} becomes 0 and, therefore, $\text{DoM}@k = 1$ applies. In contrast, if all attributes disagree, d_k^{Ham} equals the number of attributes L and results in a $\text{DoM}@k$ of 0. However, using this definition may result in overestimation of a retrieval system's performance since, particularly for queries with commonly occurring soft biometric characteristics, high DoM scores for many gallery samples might be observed. To eliminate this weakness, DoM scores are normalized w.r.t. the specific query by computing the average degree of

agreement $\overline{\text{DoM}}$ across all samples included in the gallery $\mathcal{G}$ as

$$\overline{\text{DoM}} = \frac{1}{|\mathcal{G}|} \sum_{k=1}^{|\mathcal{G}|} \text{DoM}@k. \tag{4.9}$$

Then, the normalized DoM is calculated similar to min-max normalization, but with $\overline{\text{DoM}}$ instead of the minimum value:

$$\text{DoM}_{\text{Norm}}@k = \max\{0, \frac{\text{DoM}@k - \overline{\text{DoM}}}{1 - \overline{\text{DoM}}}\}. \tag{4.10}$$

As it is assumed that each query has at least one match in the gallery (see Section 4.3), the maximum value of DoM@k is always 1. To avoid negative values, the maximum of 0 and the resulting score is computed. The normalization procedure ensures that gallery samples matching the query with the average number of attributes across all gallery samples or less result in zero precision. Afterward, the final score is calculated analogous to AP and mAP but with the new definition of precision $\text{Prec}_{\text{DoM}}@k$.

In addition to the avoidance of inconclusively high scores, normalizing the values concerning the query has another advantage. The comparability of resulting scores across different queries and also entire datasets is enhanced. Typically, the absolute values achieved for mAP but also R-1 strongly depend on the gallery set and corresponding queries. For instance, simple queries with many positive samples in the gallery set typically receive higher scores than queries searching for rarely occurring combinations of attributes. The proposed normalization procedure alleviates this by adapting the DoM computation to the respective query. As a result, achieving high DoM scores get more difficult for queries with common attributes. This reflects the fact that in a real-world scenario it is most crucial to distinguish relevant and irrelevant results. If many persons are similar to the query, small divergences concerning semantic attributes might be decisive. Conversely, if most of the people in the gallery clearly do not match the query attributes, persons with different characteristics might be interesting even if they do not match the query in its entirety. As a result, the mADM metric is a better indicator of the usefulness and quality of results in real-world application than mAP or R-1.

4.3 Evaluation Protocols

In this section, the evaluation protocols for both the PAR and attribute-based person retrieval tasks are detailed. This thesis categorizes evaluations into two distinct cases: specialization and generalization.

The specialization case involves conducting experiments on individual datasets, where training and test data originate from the same dataset. On the other hand, UPAR enables assessing the generalization capability of PAR and attribute-based person retrieval methods. The proposed evaluation protocols use different datasets for training and testing, employing a cross-validation scheme to ensure comprehensive assessment.

4.3.1 Specialization

For the PAR task, evaluation measures are computed for all test images. In the case of person retrieval, the procedure is as followed. One attribute query is created for each unique set of binary soft biometric attributes present in the test sets of the datasets. Consequently, each test image matches exactly one query. The other way round, multiple test images may match the same query if they depict persons with identical attributes. Additionally, all attributes are treated as binary, utilizing the attributes that were also evaluated in the original works.

A summary of the gallery statistics, including the number of binary attributes, the size of the gallery, and the number of attribute queries is provided in Table 4.3. The table demonstrates that a larger gallery does not necessarily lead to an increased number of distinct sets of semantic attributes. For instance, the Market-1501 dataset contains the largest gallery but the fewest number of attribute queries. The reason is the comparatively low number of attribute annotations. The PETA dataset includes the smallest gallery, but the second most queries. Due to the large number of annotated soft biometrics, the gallery set of the RAPv2 encompasses the clearly highest number of distinct attribute sets.

Table 4.3: Evaluation statistics – The table illustrates the properties of the retrieval galleries of the four specialization datasets. †The number of gallery images refers to the number of tracks.

Dataset	Bin. attributes	Gallery images	Queries
PETA	35	7,600	2,242
Market-1501	30	19,732	484
PA-100K	26	10,000	849
RAPv2	54	16,985	9,350
MARS†	52	6,848	1,949

4.3.2 Generalization

The individual datasets do not include significant domain shifts as present in real-world applications. The UPAR dataset was proposed to close this gap and, thus, offers the possibility to pursue different evaluation protocols. These protocols follow the concept of domain generalization [Bla11]. Two different protocols are proposed to assess the generalization ability of models which discriminate themselves concerning the amount of data that is available for training. Both protocols are based on cross-validation since there are four different sub-domains. The first one, referred to as 4-Fold Cross-Validation (4FCV), is the more complicated challenge since only data from a single dataset may be used during the training phase. The second evaluation scheme is Leave-One-Out Cross-Validation (LOOCV). It assumes that diverse training data from multiple sources is available. Only one sub-domain is left out for evaluation in each of the four folds and the other three are used for training. The breakdown of the datasets among the splits is shown in Table 4.4.

Table 4.4: UPAR splits – Split definitions for the two UPAR generalization evaluation schemes. The 4FCV protocol is more challenging since only data from a single sub-dataset is used for training. In contrast, the LOOCV protocols allows using data from multiple domains for training. In both cases, evaluation is performed on unseen domains.

Split	LOOCV		4FCV	
	Training	Evaluation	Training	Evaluation
1	PA-100K, PETA, RAPv2	Market-1501	Market-1501	PA-100K, PETA, RAPv2
2	Market-1501, PETA, RAPv2	PA-100K	PA-100K	Market-1501, PETA, RAPv2
3	Market-1501, PA-100K, RAPv2	PETA	PETA	Market-1501, PA-100K, RAPv2
4	Market-1501, PA-100K, PETA	RAPv2	RAPv2	Market-1501, PA-100K, PETA

Moreover, details on the number of training, validation, and test images, as well as the number of attribute queries per split are provided in Table 4.5. These splits adhere to the original splits of the single datasets to enable comparability of results.

Table 4.5: UPAR split statistics – Statistics of the four splits for each of the two evaluation protocols. Attributes without a positive sample in the training set are excluded which is why split 1 and 3 of the 4FCV scheme evaluate less than 40 attributes.

Split	LOOCV				
	Bin. attributes	Train. images	Val. images	Test images	Queries
1	40	135,124	27,465	13,093	821
2	40	69,047	20,908	9,986	1,479
3	40	139,380	29,096	6,963	1,763
4	40	100,593	15,021	15,817	3,167
Split	4FCV				
	Bin. attributes	Train. images	Val. images	Test images	Queries
1	35	12,924	3,365	32,766	4,855
2	40	79,001	9,922	35,873	5,081
3	39	8,668	1,734	38,896	4,834
4	40	47,455	15,809	30,042	3,356

It is important to note that the number of images does not exactly match the counts in the sub-datasets. During the creation of UPAR, images with inconsistent annotations or those lacking the depiction of humans were removed. Furthermore, attributes without a positive example in the training set are excluded from evaluation, as learning to recognize these soft biometric characteristics becomes impossible. This exclusion affects two folds of the 4FCV scheme, specifically split 1 and 3.

The final evaluation scores for both protocols are computed in the following manner. Metrics are first calculated for each test domain independently. For the 4FCV protocol, then, the average across test domains is computed to obtain per split results. The mean over the splits yields the final result for both protocols. This procedure ensures that each evaluation subset has an equal

influence on the final result, preventing sub-domains with large test sets and a high number of attribute queries from prevailing the others.

5 Baseline

This chapter introduces the fundamental principle of PAR as the feature extraction approach for attribute-based person retrieval. The chapter begins with a formal problem description, outlining the inputs, outputs, and goals of PAR and attribute-based person retrieval in Section 5.1. Furthermore, a multi-label classification method inspired by Jia et al. [Jia21b] is described to establish a baseline for this thesis. Details are presented in Section 5.2.

5.1 Problem Formulation

Consider a dataset $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{y}_i) \mid i = 1, 2, \dots, M\}$ consisting of M person images with corresponding labels for L soft biometric attributes. $\mathbf{I}_i$ represents the i-th image in the dataset and $\mathbf{y}_i \in \{0, 1\}^L$ denotes the binary label vector. Each element y_{ij} in this vector either indicates the presence or absence of the specific attribute. So, y_{ij} stands for the annotation of the j-th attribute for the i-th image. If $y_{ij} = 1$ applies, $\mathbf{I}_i$ is annotated to show a person with the j-th attribute and $y_{ij} = 0$ if not. PAR aims at correctly recognizing these semantic attributes for unseen test images. Typically, deep neural networks are applied for this task nowadays. With $\mathbf{x}_i$ being the network's output vector for $\mathbf{I}_i$, x_{ij} represents the specific output for the j-th attribute accordingly. Then, model outputs are transformed into so-called confidence scores through the application of an activation function. The resulting prediction vector $\mathbf{p}_i$ contains a score for each attribute that is interpreted as probability of the presence of the attribute in the input image. Usually, the sigmoid function is applied to compute $\mathbf{p}_i$ based on $\mathbf{x}_i$. Finally, output probabilities are converted to binary attribute predictions by using a threshold vector $\mathbf{t} \in [0,1]^L$. Typically, the thresholds for all attributes are set to 0.5.

The goal of attribute-based person retrieval is to sort G images $\mathbf{I}_i$ included in a huge gallery database $\mathcal{G} = \{\mathbf{I}_i | i = 1,2,\dots,G\}$ according to their distance to a person description. In this work, such descriptions are encoded as binary query vectors $\mathbf{q} \in \{0,1\}^N$, similar to the attribute labels. N denotes the number of query attributes. Each element in $\mathbf{q}$ represents whether a person with or without a specific soft-biometric characteristic is searched. Retrieval queries may include either entire person descriptions using all predictable attributes by the classifier ($N = L$) or only a certain subset ($N < L$). Due to simpler notation, the first case is assumed in the following that the query contains information for all attributes in each case. Typically, the Euclidean distance function between the binary query vector $\mathbf{q}$ and attribute prediction vectors $\mathbf{p}_i$ is calculated to determine the similarity between queries and gallery samples. By sorting the gallery samples in ascending order according to their distance from the query, the final retrieval ranking is constructed.

5.2 Strong Baseline for Pedestrian Attribute Recognition

The base framework for all the experiments conducted in this thesis is the work of Jia et al. [Jia21b]. The authors carried out a detailed study on several important aspects of PAR models and achieve results comparable to the current state-of-the-art in terms of this task. Since PAR is a multi-label classification task, the baseline model builds on a normal classification architecture, which is depicted in Figure 5.1. It consists of a CNN as the backbone, depicted in orange, which is used to extract feature maps, given input images $\mathbf{I}_i$. Subsequently, the spatial dimensions are reduced by a pooling operation to obtain a feature vector for the input image. This vector is then passed to a classification head, highlighted in blue, which computes the attribute predictions. First, a FC layer is employed to reduce the global feature size to the number of soft biometrics that should be recognized. The resulting attribute logits included in $\mathbf{x}_i$ are then projected into confidence scores $\mathbf{p}_i$ for the presence of the attributes using the sigmoid function. In the following, the feature extraction function is referred to as $\mathcal{F}(\cdot;\theta_\mathrm{f})$ and the classifier function as $\mathcal{C}(\cdot;\theta_\mathrm{c})$ with

θ_f and θ_c being the learnable parameters of the backbone and the classifier, respectively.

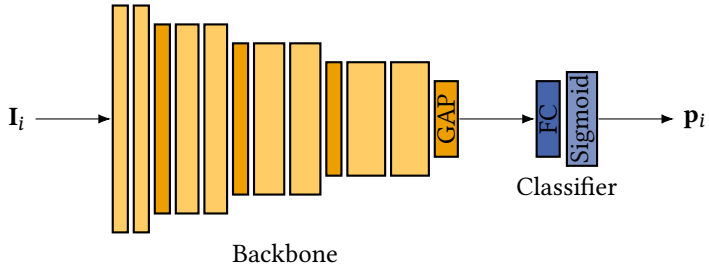


Figure 5.1: Baseline architecture – The backbone extracts feature maps for input images $\mathbf{I}_i$. Subsequently, a pooling operation, Global Average Pooling (GAP) in this case, reduces the spatial dimensions to obtain feature vectors. The FC layer serves as classification layer and consists of as many outputs as there are binary attributes to recognize. Finally, the sigmoid activation function is applied to transform the logits into values that are interpreted as probabilities for the presence of the attributes.

5.2.1 Backbone

The task of the backbone model is to extract feature maps, *e.g.*, using convolutions. It typically contains multiple stages to reduce the spatial resolution and to increase the number of feature channels. With increasing depth in the network, the task-specific semantics of features increase while low-level information gets less important. Residual Net (ResNet) [He16] architectures and several derivatives, such as ResNeXt [Xie17], Res2Net [Gao21], or ResNeSt [Zha22], established themselves the standard choice for many years as they offer state-of-the-art performance in a variety of different tasks at reasonable computational costs [Jeo21, Ji22, Ye22]. In contrast to previous CNN architectures such as VGG [Sim14], ResNet models allow deeper networks since He et al. [He16] solved the so-called degradation problem [Sri15, He15]. The degradation issue describes the phenomenon where, despite the increase in model capacity with deeper architectures, the model's accuracy saturates or even decreases at a certain point instead of improving as expected. The problem was solved by introducing residual skip connections which allow the model to only learn complementary information to the input in each block.

Typically, the parameters of the backbone model are initialized with pre-trained weights obtained by training the model on the ImageNet dataset with either 1,000 [Rus15] or 21,000 [Kol20] different classes. Thereby, the backbone is already able to produce meaningful features and less data is required for fine-tuning the model on the downstream task, *i.e.*, PAR in this case.

To reduce the feature maps into a feature vector for each image, spatial pooling is applied. The most prominent examples are Global Average Pooling (GAP) and Global Maximum Pooling (GMP) which compute the average or maximum value for each spatial feature map. The dimension of the output vector then equals the number of channels, *i.e.*, the number of filter kernels from the preceding convolution layer. For their strong baseline, Jia et al. [Jia21b] apply GAP. Since a detailed analysis [Spe20b] showed only negligible impact of the pooling function on the model's performance, the same procedure is followed.

5.2.2 Classifier

The classifier module maps the feature vectors to the output dimension, *i.e.*, the number of binary semantic attributes. Jia et al. [Jia21b] rely on the straightforward approach to use a single FC layer. Each input neuron is connected to each output neuron so that all of the attribute predictions learn independent weights for the importance of different input dimensions.

Thereafter, output logits x_{ij} are activated by the sigmoid function as follows:

$$p_{ij} = \frac{1}{1 + e^{x_{ij}}}. \tag{5.1}$$

As a result, values for p_{ij} are normalized to fall into the interval [0,1] and, therefore, are interpreted as probabilities for the presence of attributes.

5.2.3 Loss Function

The task is dealt with as a multi-label classification task. Consequently, the overall objective is to learn the parameters θ_f and θ_c of the feature extractor and classifier by minimizing the empirical risk loss:

$$Loss = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_{\text{PAR}}(\mathbf{y}_i, \mathcal{C}(\mathcal{F}(\mathbf{I}_i; \theta_f); \theta_c)). \tag{5.2}$$

Commonly, the weighted binary cross-entropy loss is chosen as the loss function $\mathcal{L}_{\text{PAR}}$ in PAR [Sch18, Jia21b, Spe23b], which is formulated as

$$\mathcal{L}_{\text{CE}} = -\sum_{j=1}^{L} w_j [y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij})]. \tag{5.3}$$

Attribute weights w_j are vital to mitigate the influence of imbalanced attribute distributions, which pose a major challenge in classifying rarely occurring soft biometrics. The attribute-specific weights ensure that training samples depicting such attributes receive enhanced focus by increased loss values. However, the diversity of appearances of the attributes with few positive samples stays limited. Consequently, this technique carries the risk of overfitting, particularly when applied excessively. Multiple functions were proposed in literature to compute these weights, all of which utilize the positive ratio of attributes in the training set [Li15, Tan20, Zha21b]. This ratio r_j^{pos} represents the proportion of training images showing the j-th attribute. However, Jia et al. [Jia21b] show that the influence between different weighting functions is negligible and that it is only important to use such weights to handle imbalanced distributions. The higher the weights the better the recall of positive samples and, therefore, the label-based mA score. Concerning the instance-based F1 (F1), the recall increases to the detriment of the precision. Based on these findings, the weight function proposed by Li et al. [Li15] was chosen for the

experiments in this thesis. It is defined as

$$w_j = \begin{cases} e^{1-r_j^{\text{pos}}} & \text{if } y_{ij} = 1 \\ e^{r_j^{\text{pos}}} & \text{if } y_{ij} = 0 \end{cases}. \tag{5.4}$$

5.2.4 Implementation Details

The methodology introduced in this thesis is implemented using the *PyTorch*[1] deep learning framework. In terms of hyperparameters, the values chosen for training the baseline closely follow those proposed by Jia et al. [Jia21b].

For model selection, evaluation is done on the validation sets of the datasets. The number of epochs for which models are trained varies depending on the specific dataset, considering the diverse characteristics and sizes of the research datasets. PAR models tend to quickly overfit due to limited diversity in training datasets. Therefore, training on the datasets with a fixed number of epochs is avoided to prevent overfitting as well as underfitting.

An initial learning rate of $1e^{-4}$ is chosen with Adam as the optimizer. The learning rate is reduced by a factor of 0.1 when the results measured by the validation loss do not improve for 4 subsequent epochs. The weight decay parameter is set to $5e^{-4}$ for all experiments.

Regarding initialization, the learnable parameters of the backbone model are initialized using pretrained weights from the ImageNet dataset [Rus15]. The classifier and additionally added layers are initialized randomly.

Furthermore, Jia et al. [Jia21b] explored the influence of different input image sizes on PAR results. Following their findings, input person images are scaled to be 192 pixels in width and 256 pixels in height. Batches consist of 64 person images. Random horizontal flipping and random cropping are applied as data augmentation techniques to increase the diversity in the training datasets.

[1] https://pytorch.org/

During evaluation, recognition thresholds for all attribute are set to 0.5. For generating the retrieval rank lists, the default distance measure employed is the Euclidean distance.

6 Pedestrian Attribute Recognition

The concept of this thesis is based on PAR models as semantic feature extractors to enable attribute-based person retrieval. The focus is on global image-based approaches since these models are computationally efficient, making them advantageous for real-world applications. Fast inference of of individual components is crucial, particularly if PAR represents just one part of a comprehensive framework. Furthermore, Gkioxari et al. [Gki15] discovered that with the emergence of deeper networks, the difference in accuracy to more complex models, such as part-based models, diminishes. This suggests that the observed improvements are mainly due to the increased capacity of models and employing current backbone architectures in global models may accomplish comparable performance. Furthermore, Jia et al. [Jia21b] demonstrate that modern global models are capable of achieving strong performance.

Due to the specific challenges raised by real-world surveillance systems (see Section 1.2), several adaptations are made to the baseline which are elaborated on in this chapter. First, an in-depth examination of PAR model design choices and their impact on attribute-based person retrieval is carried out in Section 6.1. Additionally, in Section 6.2, the use of normalization modules is thoroughly studied to balance PAR results and simultaneously enhance attribute-based person retrieval. Finally, several strategies are explored in Section 6.3 to recognize soft biometrics based on tracks of individuals rather than single images, allowing for video-based processing.

6.1 Evaluation of Design Choices

This section explores various design choices concerning PAR and evaluates their impact on accuracy and generalization capability. The main idea is to optimize the baseline architecture and the training process presented in Section 5.2. Two primary reasons motivate this approach. First, simple architectures provide faster inference, which is vital for real-world deployment. Second, research suggests that complex models with many learnable parameters often underperform simpler models regarding generalization due to overfitting to the training data [Cor23, Spe23b].

This thesis differs from existing studies, especially Jia et al. [Jia21b], by focusing on the task of attribute-based person retrieval. Instead of solely optimizing the intermediary task of PAR, the influence of design choices on attribute-based person retrieval is considered for the first time. The author of this thesis published several papers [Spe20c, Cor23, Spe23b], highlighting that PAR metrics may not be dependable indicators for the quality of attribute-based person retrieval. This is because thresholds are applied in PAR to obtain binary predictions, which are then employed for calculating the metrics, diminishing the importance of actual confidence scores. In contrast, biased output probabilities caused by over- or underconfident models affect the computation of the Euclidean distances used to create the retrieval rankings.

The research outlined in this section is primarily based on three publications by the author [Sch18, Spe20b, Spe23b].

6.1.1 Binary vs. Multi-Class Attributes

In this thesis, all attributes are treated as binary and, thus, the sigmoid activation function is applied to transform the model's output logits into values that are understood as prediction probabilities for the presence of attributes. Nonetheless, binary attributes and multi-class attributes could conceivably be differentiated and, therefore, an argument could be made for such as distinction. Unlike binary attributes that only differentiate between the presence and absence of soft biometric characteristics, multi-class attributes have more than

two manifestations. These attributes are labeled using the one-hot encoding technique, where only one manifestation is possible per image. An example of a multi-class attribute is age. Since estimating a person's actual age from surveillance imagery is challenging, age is commonly categorized into distinct classes, such as *child*, *teenager*, *adult*, or *elderly*. Using the softmax activation function for these attributes forces the model to acknowledge correlations between manifestations and learn that only one attribute can be correct for each image. The softmax function generates a probability distribution for the manifestations, rather than producing independent confidence scores.

Comparative experiments were conducted to examine the decision to use the pure binary approach. The outcomes of these experiments are exhibited in Table 6.1.

Table 6.1: Binary vs. multi-class attributes – Treating all attributes as binary outperforms the combined approach concerning instance-based F1 and person retrieval metrics. In contrast, using the softmax activation for multi-class attributes improves the recognition of individual attributes, as stronger results for label-based mA are achieved.

Attributes	Market-1501				
	mA	F1	mADM	mAP	R-1
Binary	76.4	**85.2**	**60.4**	**25.5**	**37.8**
Binary & multi-class	**78.7**	80.6	58.4	22.1	32.6

It is advantageous to consider all attributes as binary and independent, as observable from the results. Only the mA benefits from utilizing multi-class attributes. In the multi-class case, the softmax activation function increases the recall of positive samples, particularly for uncommon attribute manifestations. For instance, the positive recall of yellow lower-body clothing is enhanced from 8.3% to 50.0%. As the model is compelled to produce a high probability value for the most probable manifestation and zero for the rest, predictions for rare attributes surpass the attribute threshold more frequently. However, calibration of the outputs deteriorates, and there is an elevated risk of the model becoming overconfident in its predictions [Mül19]. This leads to a loss of correlation between the model's actual certainty and the generated confidence scores. As a result, treating all attributes as binary is clearly

preferable, given this negative impact. The independence of all attributes allows for more information to be encoded in the attribute prediction vectors, leading to improved discrimination between individuals during distance computation for creating the retrieval result.

As a result of these findings, this thesis adopts the all binary approach, including different exclusive categories such as age. The models appear capable of implicitly considering and learning the correlations among binary attributes sufficiently.

6.1.2 Backbone

Many research methods in the PAR domain continue to rely on ResNet [He16] in the variant ResNet-50 as the backbone model due to its competitive performance and low computational costs on various tasks. Nevertheless, there are several recently proposed network architectures that surpass this model in various tasks [Gao21, Liu21, Wan21, Liu22, Wan22b]. On the one hand, transformer-based models gain increasing importance. Popular transformer backbones for vision tasks include the Vision Transformer (ViT) [Dos21], Swin [Liu21], and Pyramid Vision Transformer v2 (PVTv2) [Wan22b]. On the other hand, novel CNNs architectures with improved performance have been developed, such as ConvNeXt [Liu22]. Since, besides the retrieval accuracy, inference time is a crucial measurement to assess the suitability for application in the real world, Figure 6.1 provides a comparison of various backbone architectures' mADM and corresponding inference times. The results were obtained on the PETA dataset using the baseline approach described in Section 5.2. Inference times are averaged over the test split for processing a single image in each step. As GPU, an *NVIDIA GeForce RTX 3090* was employed. The plot points out the advantages of the ConvNeXt and Swin architectures compared to their CNN and transformer counterparts ResNet and PVTv2, respectively. Without regard to the model variant, both achieve decisively stronger outcomes. When comparing those two top-performing architectures, it is discernible that the largest version of the Swin transformer

achieves the best overall results, however, with a considerably slower inference speed than the CNN ConvNeXt. In general, transformer-based models exhibit longer inference times than those observed for CNNs, which achieve similar scores for mADM.

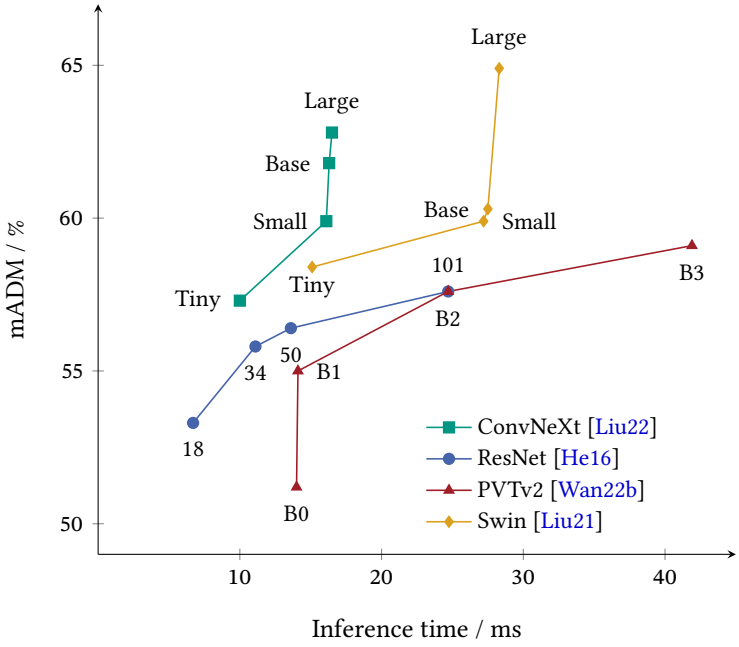


Figure 6.1: Comparison of backbones architectures – The plot visualizes the influence of various backbone architectures on attribute-based person retrieval results and inference times. The ConvNeXt and Swin architectures clearly outperform ResNet and PVTv2.

Another interesting finding is the slight increase in inference time between the Base and Large variants of the ConvNeXt and Swin architectures. The curves for both model architectures are similar since ConvNeXt was developed by transforming network design elements and structure from the Swin transformer into the CNN world. Larger variants mainly expand in width, which can be efficiently computed in parallel by modern GPUs. However,

the number of parameters and, thus, the memory footprint significantly increase, requiring more expensive GPUs with larger memory. For instance, the ConvNeXt-Large model has more than twice as many parameters than ConvNeXt-Base [Liu22]. Furthermore, the plot clearly indicates that the popular ResNet architecture no longer achieves state-of-the-art results. More recent CNNs as well as transformer backbones outperform ResNet with a similar or reasonable increase in inference time.

These findings are valid for further specialization datasets and metrics as well, as shown in Table 6.2.

Table 6.2: Specialization results for different backbones – Using Swin-Large as the backbone model consistently leads to the best results on the datasets. Concerning CNN architectures, ConvNeXt variants clearly outperform the ResNet-50 model.

Backbone	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
ResNet-50	84.2	86.4	56.4	20.5	20.7	82.1	88.3	67.3	24.3	33.5
ConvNeXt-Base	86.1	88.1	61.8	24.4	24.4	82.2	88.5	68.8	26.2	34.5
ConvNeXt-Large	86.6	88.5	62.8	25.6	25.6	83.2	89.4	71.6	28.7	35.7
PVTv2-B2	84.4	87.1	57.6	21.6	22.0	81.9	88.9	68.9	26.6	35.7
Swin-Base	86.5	87.6	59.9	22.8	22.3	83.5	88.1	67.9	24.9	31.3
Swin-Large	**88.0**	**89.2**	**64.9**	**28.2**	**28.3**	**84.5**	**89.5**	**72.3**	**29.8**	**37.6**
Backbone	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
ResNet-50	77.5	78.6	54.6	17.5	12.0	76.4	85.2	60.4	25.5	37.8
ConvNeXt-Base	79.3	80.0	58.5	20.5	14.5	**80.7**	85.7	65.9	31.6	**47.7**
ConvNeXt-Large	79.0	80.1	58.2	20.3	14.3	79.8	84.7	65.3	30.3	45.7
PVTv2-B2	78.1	79.5	56.4	19.0	13.4	77.4	83.7	60.8	24.5	36.6
Swin-Base	79.5	79.8	57.0	19.2	13.4	77.3	84.4	62.4	26.7	39.9
Swin-Large	**80.4**	**80.2**	**58.6**	**21.0**	**15.2**	**80.7**	**85.9**	**69.0**	**34.3**	47.1

The highest performance across all datasets for the specialization scenario is achieved by Swin-Large, followed by ConvNeXt-Large on the PETA and PA-100K dataset, and ConvNeXt-Base on the RAPv2 and Market-1501 dataset, respectively.

However, a different picture emerges when examining the outcomes for the cross-domain generalization settings on the UPAR dataset. According to the experimental evaluations in Table 6.3, ConvNeXt-Large is prone to overfitting

and, thus, yields inferior results than the more compact Base variant. Furthermore, the advance of Swin-Large is less substantial in the generalization case. In the attribute-based retrieval task, using ConvNeXt-Base as backbone even outperforms the utilization of the Swin transformer, as per the LOOCV evaluation protocol which uses multiple data sources for training. Therefore, the resulting difference in performance does not justify the increased memory requirements and inference time for the Swin-Large backbone.

Table 6.3: Generalization results for different backbones – Contrary to the specialization scenario, using the ConvNeXt-Base model as backbone is superior to the utilization of the larger variant. Moreover, the advantage of the Swin-Large backbone vanishes. In terms of the LOOCV evaluation protocol, leveraging ConvNeXt-Base as the backbone even outperforms the performance achieved with Swin-Large.

Backbone	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
ResNet-50	71.1±2.0	78.7±3.0	49.0±7.1	13.3±4.6	16.4±10.0
ConvNeXt-Base	73.2±1.9	82.1±2.7	**56.1**±5.1	**17.6**±4.3	**19.2**±8.3
ConvNeXt-Large	**74.3**±2.1	81.4±2.6	55.2±5.1	16.8±4.4	18.7±9.4
PVTv2-B2	72.4±1.4	80.8±2.8	52.5±6.0	15.2±4.6	16.9±8.6
Swin-Base	72.7±2.7	80.2±2.6	50.8±5.3	13.9±4.1	16.3±9.1
Swin-Large	73.7±1.7	**82.2**±2.2	55.3±4.8	16.9±4.6	18.8±9.2
Backbone	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
ResNet-50	65.4±2.4	71.1±5.9	36.4±6.5	6.5±3.3	8.2±4.6
ConvNeXt-Base	**70.1**±1.5	77.2±4.2	46.4±5.8	11.0±3.6	12.8±4.6
ConvNeXt-Large	68.8±1.8	76.0±5.1	44.8±6.6	10.1±4.1	11.4±4.8
PVTv2-B2	68.2±1.7	74.7±4.4	41.7±6.0	8.4±3.3	10.0±4.2
Swin-Base	68.7±1.9	73.4±6.0	40.4±6.6	8.2±3.7	10.1±4.8
Swin-Large	70.0±1.7	**77.8**±3.5	**46.9**±5.1	**11.4**±3.2	**12.9**±4.5

In conclusion, this thesis focuses on using ConvNeXt-Base as the backbone model due to its favorable generalization performance and tradeoff between accuracy, inference time, and memory footprint.

6.1.3 Stochastic Weight Averaging

Averaging a model's weights over multiple iterations during the training process is a technique known as Stochastic Weight Averaging (SWA). Izmailov et al. [Izm18] demonstrate that SWA leads to flatter minima in the training loss, enhancing performance especially concerning generalization. Besides, in the context of PAR, trainings are prone to instability, resulting in significant variations in results between epochs, thereby aggravating the problem of selecting appropriate model snapshots for deployment.

This phenomenon is visualized in Figure 6.2. The achieved mAP values for each training epoch of the baseline model, with ConvNeXt-Base as its backbone, are presented in two ways: first, without SWA and second, with Simple Moving Average (SMA) [Arp22], as an example of SWA techniques. The training was conducted using the PETA dataset.

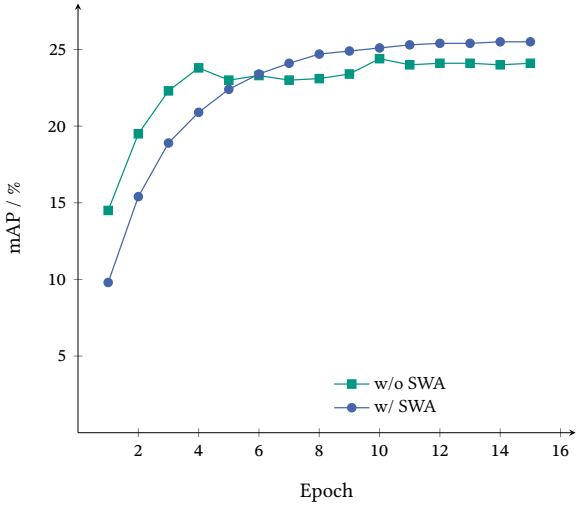


Figure 6.2: Impact of SWA on the training process – The achieved mAP values in each training epoch are illustrated. On the one hand, for the baseline approach and, on the other hand, with the use SMA [Arp22] as a SWA technique. Using SMA eliminates fluctuations and simultaneously improves the mAP.

While training without SWA leads to fluctuating mAP scores from epoch to epoch, and random peaks of performance, for instance in epoch 10, the SWA training smoothly converges toward its maximum result. In addition, the averaged model outperforms the baseline approach without SWA. As a result, choosing suitable model snapshots for deployment is made easier due to the similarity of results in neighboring epochs.

To mitigate the aforementioned issues and improve performance, particularly in terms of generalization, various variants of SWA are compared to identify the best suitable SWA technique for PAR and attribute-based person retrieval.

Arpit et al. [Arp22] propose the SMA protocol, which averages weights starting from iteration t_0 with a specified sampling frequency. With θ_t being the model's parameters at iteration t, the averaged model parameters $\hat{\theta}_t$ are calculated according to

$$\hat{\theta}_t = \begin{cases} \theta_t, & \text{if } t \leq t_0 \\ \frac{t-t_0}{t-t_0+1} \cdot \hat{\theta}_{t-1} + \frac{1}{t-t_0+1} \cdot \theta_t, & \text{otherwise} \end{cases}. \tag{6.1}$$

Empirical analysis conducted by Arpit et al. [Arp22] indicates that setting the hyperparameters t_0 and the averaging frequency close to 0 and 1, respectively, yields satisfactory results. This finding holds true in the experiments in this thesis for the PAR and attribute-based person retrieval domains. Consequently, this method is considered parameter-free.

The SMA averages the model parameters at each iteration t so that equal contribution of each set of model parameters θ_t to the final averaged model is ensured. However, since the model is successively improving over the course of training, assigning higher weights to later iterations could be beneficial. This is achieved by using the Exponential Moving Average (EMA) update rule for averaging the weights. This approach is commonly employed in various works from the literature [Tan19a, Gle22]. The averaged model at iteration t is computed as

$$\hat{\theta}_t = \begin{cases} \theta_t, & \text{if } t \leq t_0 \\ \alpha_{\text{EMA}} \cdot \hat{\theta}_{t-1} + (1 - \alpha_{\text{EMA}}) \cdot \theta_t, & \text{otherwise} \end{cases}, \tag{6.2}$$

where α_{EMA} is a hyperparameter that controls the decay. A higher value of α_{EMA} increases the influence of early iterations on the averaged model. In the experiments, $\alpha_{\text{EMA}} = 0.9998$ proved to be suitable for the tasks addressed in this thesis.

The last approach considered in the study is Stochastic Weight Averaging Densely (SWAD) [Cha21]. Similar to the SMA method, model parameters θ_t from sampled iterations contribute equally to the averaged model. However, unlike SMA, SWAD aims at automatically determining the start and end iterations t_0 and t_e for the averaging process by detecting the validation loss valley.

The differences between these SWA methods are illustrated in Figure 6.3, which demonstrates the impact of iterations on the final averaged model parameters. It is important to note that the hyperparameters were selected arbitrarily.

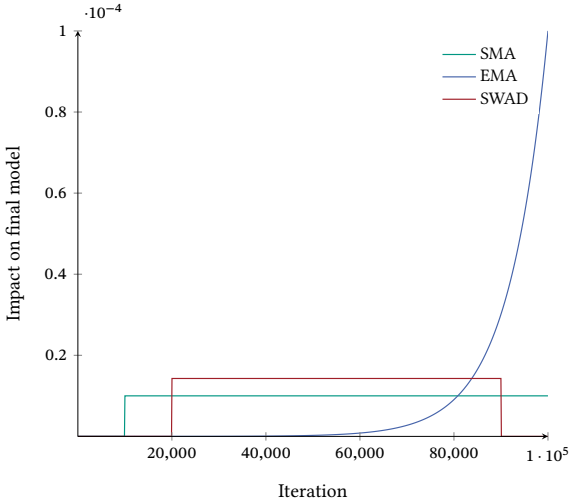


Figure 6.3: Comparison of SWA methods – The plot visualizes the impact of the model weights in each iteration on the final weights obtained through weight averaging. The EMA update rule exponentially increases the impact of iterations on the resulting model. The impact of averaged iterations is greater for SWAD than for SMA due to the smaller number of iterations included in the average model. Note that the hyperparameters to create the curves were chosen arbitrarily.

While the SMA method assigns equal impact to all iterations after the starting iteration, SWAD samples model parameters within a specific window. This means only a subset of iterations contributes to the averaged model. In contrast to both, the EMA technique increases the influence of weights for progressing training in an exponential manner.

Experimental findings indicate that automatically detecting a specific window to average the model's parameter is not favorable compared to the simple and parameter-free SMA approach. The best configuration is to start averaging from the beginning and stopping when the validation results no longer improve. In this case, SWAD delivers equivalent results to SMA, which is why SWAD is not considered further.

Table 6.4 compares the results achieved with the SWA techniques SMA and EMA on the four specialization datasets.

Table 6.4: Specialization results for SWA techniques – The use of SMA during training leads to consistent improvements on each of the datasets and regarding both PAR as well as attribute-based person retrieval. In contrast, employing EMA only achieves superior performance than the baseline on the large PA-100K and RAPv2 datasets.

SWA	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
–	86.1	88.1	61.8	24.4	24.4	82.2	88.5	68.8	26.2	34.5
SMA	**86.5**	**88.4**	**63.4**	**25.5**	**25.1**	**84.3**	**90.0**	**72.3**	**30.4**	**38.8**
EMA	82.4	86.0	60.9	22.1	22.0	83.7	89.6	71.7	29.5	36.4
SWA	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
–	79.3	80.0	58.5	20.5	14.5	**80.7**	85.7	65.9	31.6	47.7
SMA	**79.4**	**81.1**	**60.6**	**22.6**	**16.3**	79.7	**86.6**	**67.9**	**34.0**	**50.6**
EMA	78.5	80.9	60.2	22.2	16.0	77.5	86.2	64.5	29.9	45.0

The results demonstrate that SMA clearly outperforms both the baseline as well as averaging the model's learnable parameters with the EMA update rule. Moreover, it is found that the use of EMA on small datasets, such as the PETA and Market-1501 datasets, leads to deterioration in the results of both tasks. This is due to overfitting occurring in later iterations as a result of the lack

of sufficient data. Therefore, focusing on the model parameters in these iterations is not advantageous. On the larger two datasets, an improvement is evident when compared to the baseline approach without SWA. However, the performance still lags behind the results achieved with the SMA technique. In addition, EMA models require more time to converge due to their focus on later epochs. Besides, this approach necessitates selecting adequate values for the hyperparameter α_{EMA}, which are significant disadvantages.

An important reason for the use of SWA is to improve the ability to generalize to new domains. The results so far indicate that SWA is beneficial even when the training and test data originate from the same dataset. Subsequently, the generalization performance is evaluated in Table 6.5 using the UPAR dataset.

Table 6.5: Generalization results for SWA techniques – Analogous to the specialization results, the best approach is to use the SMA method. Moreover, while the use of the EMA rule results in comparable retrieval performance when the LOOCV protocol is applied, performance on the 4FCV protocol clearly lacks behind both the SMA technique and the baseline.

SWA	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
–	73.2±1.9	82.1±2.7	56.1±5.1	17.6±4.3	19.2±8.3
SMA	**74.0**±2.4	**83.4**±2.0	**58.5**±5.4	**19.5**±5.4	**21.4**±9.3
EMA	72.1±1.7	83.2±2.0	**58.5**±5.4	19.4±5.5	20.9±9.3
SWA	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
–	**70.1**±1.5	77.2±4.2	46.4±5.8	11.0±3.6	12.8±4.6
SMA	69.5±1.9	**78.4**±3.8	**47.8**±6.4	**11.8**±4.3	**13.2**±5.3
EMA	63.3±6.3	67.2±12.8	35.9±17.3	8.1±7.5	9.3±8.7

The results obtained in this study are comparable to those found for the specialization scenario. With the exception of the mA in the 4FCV protocol, which employs a single sub-dataset for training, the use of SMA exhibits the best performance. Experiments using EMA and the 4FCV evaluation protocol demonstrated inadequate progress and insufficient results within a reasonable timeframe. Convergence in cross-domain settings typically occurs in the

early training epochs for the tasks considered in this thesis. Therefore, concentrating on later epochs does not result in improvement but rather falls short of achieving strong scores. The impact of earlier epochs with stronger performance vanishes as training progresses.

In summary, the utilization of SWA effectively reduces the oscillation of evaluation results between epochs and clearly improves PAR and attribute-based person retrieval. When comparing different approaches, SMA turned out to be the most versatile approach. Additionally, it is considered parameter-free and, therefore, is easily applicable without further tuning.

6.1.4 Loss Function

The strong baseline by Jia et al. [Jia21b] utilizes the cross-entropy loss function along with a weighting function to deal with the class imbalance problem. However, an alternative approach is to use the focal loss function [Lin17], which is a weighted form of the cross-entropy loss originally introduced for object detection tasks. As pointed out in the discussion of the related literature in Section 2.2, several works rely on variants of this loss function [Sar18b, Ji20, Zhe21]. The focal loss function introduces a multiplication factor that increases the importance of misclassified hard samples while ensuring that low loss values are propagated for easy samples. As a result, the model learns to recognize semantic attributes under difficult conditions or based on few, small cues. In addition, the same loss weighting mechanism described in Section 5.2.3 can be applied to further address class imbalances. Let w_j represent this positive ratio-based loss weight for the j-th attribute. The weighted focal loss function $\mathcal{L}_{\text{FL}}$ is then defined as

$$\mathcal{L}_{\text{FL}} = -\sum_{j=1}^{L} w_j[(1-p_{ij})^{\alpha_{\text{FL}}} y_{ij} \log p_{ij} + p_{ij}^{\alpha_{\text{FL}}}(1-y_{ij}) \log(1-p_{ij})]. \quad (6.3)$$

For positive and negative samples of the j-th attribute, the focal loss weight is $(1-p)^{\alpha_{\text{FL}}}$ and $p_{ij}^{\alpha_{\text{FL}}}$, respectively, *i.e.*, the difference between prediction and ground truth values controlled by a relaxation parameter α_{FL}. The higher α_{FL}

the more importance is rewarded to misclassified training examples compared to simpler ones. The remainder of the equation follows the definition of the cross-entropy loss function provided in Section 5.2.3.

First, the influence of the hyperparameter α_{FL} is examined. The study of Lin et al. [Lin17] revealed that 2 is generally a suitable choice for the detection task. However, due to notable differences from the tasks addressed in this work, Figure 6.4 provides the mADM results for different values of α_{FL} obtained using the Market-1501 and RAPv2 datasets. The results were generated using ConvNeXt-Base as the backbone model and SMA.

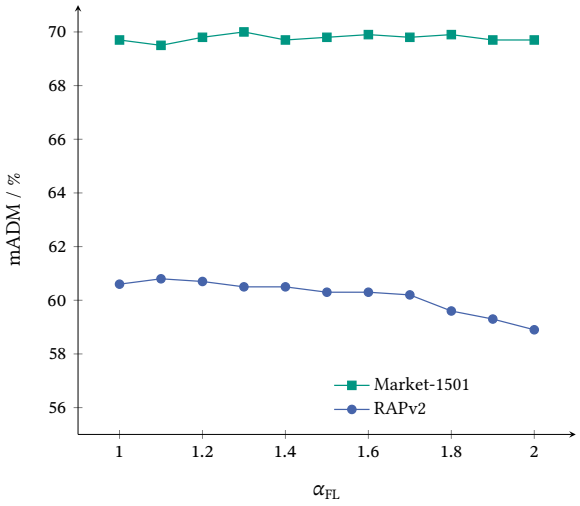


Figure 6.4: Influence of α_{FL} on the retrieval performance – While the mADM is nearly constant for increasing values of α_{FL} for the Market-1501 dataset, the mADM values decrease for the RAPv2 dataset. Therefore, α_{FL} is conservatively set to 1 for the experiments. Note that the x- and y-axis in the figure do not start at zero.

The maximum scores for mADM are reached for α_{FL} values of 1.3 and 1.1 for the Market-1501 dataset and the RAPv2 dataset, respectively. While the mADM is nearly constant for increasing values of α_{FL} when evaluated with the Market-1501 dataset, the curve drops for the RAPv2 dataset. A possible explanation is that RAPv2, due to its size, provides enough diversity to train a

robust model. Focusing too much on few challenging images, which may also be difficult to classify due to annotation errors, avoids exploiting this diversity, which in turn leads to overfitting and poor generalization performance on unseen images with different characteristics. To avoid exacerbating this negative effect, α_{FL} is cautiously set to 1 for all datasets and experiments.

Analogous to the figure, the results below are generated with ConvNeXt-Base and SMA. As can be seen in Table 6.6, the use of the focal loss function improves attribute-based person retrieval on each of the datasets except the RAPv2 dataset. In contrast, the influence of the choice of the loss function on PAR is negligible.

Table 6.6: Comparison of loss functions for specialization – Except for the RAPv2 dataset, the use of the focal loss function outperforms the cross-entropy loss concerning all retrieval metrics. The PAR results are only slightly affected by the choice of the loss function.

Loss	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Cross-entropy	**86.5**	**88.4**	63.4	25.5	25.1	**84.3**	**90.0**	72.3	**30.4**	38.8
Focal	86.4	**88.4**	**65.1**	**27.0**	**26.3**	**84.3**	89.9	**73.4**	**30.4**	**38.9**
Loss	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Cross-entropy	**79.4**	**81.1**	60.6	**22.6**	**16.3**	79.7	**86.6**	67.9	34.0	**50.6**
Focal	79.2	80.9	**61.1**	22.3	15.9	**80.1**	86.4	**69.7**	**35.5**	50.2

For the RAPv2 dataset, the increased focus on a few samples leads to biased learning of information from the training data. This is beneficial for recognizing difficult or infrequent attributes, but degrades performance on other soft biometric features that are comparatively easy to recognize. For instance, the use of the focal loss improves the recognition of difficult and rare attributes such as *body thin*, but at the same time deteriorates the mA of common attributes such as *long hair*. However, concerning retrieval, the mADM improves nevertheless.

For the other datasets, notable improvements concerning the retrieval task are achieved by the focal loss. Smaller datasets benefit more than larger datasets,

since they contain less diversity and, therefore, it is more important to focus on challenging samples.

The effect of substituting the cross-entropy loss with the focal loss on generalization is examined in Table 6.7.

Table 6.7: Comparison of loss functions for specialization – The results regarding the generalization evaluation protocols of the UPAR dataset confirm the observations made in the specialization experiments. The differences in the PAR results are minimal, while the person retrieval task benefits from the use of the focal loss function.

Loss	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
Cross-entropy	74.0±2.4	**83.4**±2.0	58.5±5.4	19.5±5.4	21.4±9.3
Focal	**74.2**±2.4	83.2±2.1	**60.1**±5.1	**20.0**±5.3	**22.0**±9.1
Loss	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
Cross-entropy	**69.5**±1.9	**78.4**±3.8	47.8±6.4	11.8±4.3	13.2±5.3
Focal	**69.5**±2.0	78.3±3.7	**49.3**±6.6	**12.4**±4.4	**13.7**±5.1

Similar to the specialization results, the focal loss mainly improves retrieval performance. The impact on the PAR evaluation metrics is negligible. This finding indicates that the expanded focus on challenging samples introduced by the focal loss weighting induces a better calibration of the resulting confidence scores, which thereby correlate more closely with the empirical probabilities. Thus, more reliable estimates are available for the calculation of the retrieval distances.

In summary, the focal loss proves superior to the cross-entropy loss for the attribute-based person retrieval task. Across all datasets, the strongest results in mADM are achieved with this loss function. The increased focus on challenging training examples based on the predicted confidence scores also performs well in generalization settings.

6.1.5 Batch Size

Choosing an adequate batch size is an important hyperparameter choice in deep learning. It defines the number of training samples that are used in each forward and backward pass during training. A large batch size typically results in increased computational efficiency and, thus, shorter training times but also may lead to reduced accuracy and overfitting. The smaller the batch size the greater the impact of each individual training image and, therefore, the ability of the model to capture fine-grained nuances in the data. Of course, if the batch size gets too small, training behavior might be unstable and strong fluctuations of performance might be observed.

Smith et al. [Smi17] found that there is direct connection between batch size and learning rate. The experiments indicate that similar effects can be achieved by adapting the learning rate or the batch size by the same factor. For instance, this would allow to achieve similar performance with larger batch sizes and, hence, faster training through adjusting the learning rate, thereby omitting the biggest disadvantage of smaller batch sizes without negative effects on accuracy. However, further works [Kes16, Hof17] show that small batch sizes are superior to larger ones in generalization tasks. According to Keskar et al. [Kes16], using large batch sizes in methods may cause them to get stuck in local minima. On the other hand, smaller batches with more diversity in batch updates tend to push out of local minima and have a bigger chance of finding the global minimum.

Since this thesis focuses on achieving strong generalization capability, it is decided to rely on adapting the batch size instead of varying the learning rate following Smith et al. [Smi17]. As an alternative, it would be possible to adjust the training scheme, since findings of Hoffer et al. [Hof17] indicate that not a reduced batch size as such leads to improved generalization but the increased number of model updates.

The current literature on PAR mostly overlooks the impact of different batch sizes on the training and the resulting model. However, especially when dealing with unbalanced attribute classification, choosing an appropriate batch size for training can play a major role due to the explanations above. If the

batch size is too large, person images depicting semantic attributes that are rarely present will have minor influence on the batch updates and, thus, on the model parameters, since they will be outweighed by samples without the attribute.

The results in Table 6.8 are consistent with the above assumptions and findings from the related literature. The results were obtained using the ConvNeXt-Base backbone with SMA and the focal loss function. B represents the batch size. To train the baseline, 64 training images form a batch.

Table 6.8: Specialization results for varying batch sizes – Datasets with fewer images, such as the PETA or Market-1501 datasets, benefit from batch sizes smaller than 64, while the results deteriorate for the larger PA-100K and RAPv2 datasets with decreasing batch sizes.

B	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
128	85.7	87.8	64.0	25.3	24.6	83.9	89.8	73.0	29.9	37.2
64	86.4	88.4	65.1	27.0	**26.3**	**84.3**	**89.9**	**73.4**	**30.4**	**38.9**
32	**86.8**	**88.6**	66.0	**27.3**	25.4	83.8	89.6	72.6	29.0	36.0
16	**86.8**	88.5	**66.3**	26.9	26.1	83.1	89.0	71.1	26.6	34.0
8	85.5	87.4	63.4	23.5	22.8	80.8	88.2	69.1	24.3	31.7
B	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
128	**79.4**	80.8	60.5	21.5	15.1	79.7	86.3	68.8	34.7	49.8
64	79.2	**80.9**	**61.1**	**22.3**	**15.9**	80.1	86.4	69.7	35.5	50.2
32	77.8	80.4	59.1	20.3	14.3	80.5	**86.7**	70.3	36.5	**51.9**
16	76.0	79.5	56.1	17.6	12.1	80.9	86.6	**70.9**	37.0	50.0
8	73.4	78.3	51.9	13.8	8.8	**81.1**	**86.7**	**70.9**	**37.2**	50.2

For small datasets, such as the PETA and Market-1501 datasets, the results of PAR and attribute-based person retrieval improve when the batch size is reduced to, *e.g.*, 32 or 16. Due to the limited diversity provided by the small number of training images, larger batch sizes lead to overfitting and prevent fine details from being captured by the model. In particular, models trained on Market-1501 benefit from reduced batch sizes. For this dataset, optimal results are obtained for mini-batches consisting of only 8 person images. This is due

to strong attribute imbalances and a small number of different individuals and, thus, outfits included in the dataset. Since the dataset was originally designed as a person re-identification dataset, only 750 persons appear in the training data, but with multiple images each. As a result, the variance in attribute combinations is limited, and some attributes are only present in one form. Furthermore, if the batch size is too small, for instance, 8 for the PETA dataset, the performance deteriorates. In this case, the models struggle to capture the overall concepts behind the attributes due to the strong influence of individual images.

However, when looking at the larger datasets, there is a deterioration in the results for batch sizes smaller than 64. The reason is that due to the size of the datasets, enough and diverse samples are available for the attributes. Thus, exploiting the regularization capabilities of larger batch sizes improves the capture of the concepts behind the attributes. For instance, since the PAR model sees many different bags during training, it is beneficial for the generalization ability to novel types of bags not to focus too much on the appearance of a particular bag. Instead, it should capture relevant features that make up a bag and that are common to the bags in the training data. Increasing the batch size to 128 leads to symptoms of overfitting for PA-100K, while negligible improvements are observed in mA for the RAPv2 dataset.

The results so far indicate that reducing the batch size and thereby increasing the number of model updates improves the accuracy for small datasets with less than 20,000 training images for both tasks considered in this thesis. In particular, datasets such as Market-1501 with low intra-class variance benefit. Larger datasets with more samples per attribute should be trained with larger batches to achieve optimal performance. General takeaways from the analysis are that training on large PAR datasets should be performed with 64 images within a batch, while a lower batch size of, *e.g.*, 16 is superior when dealing with less training data and, hence, typically limited diversity. Best results for the mADM metric are achieved with these batch size configurations.

The following Table 6.9 illustrates the findings by means of positive recall of different attributes. The positive recall describes the proportion of samples with a certain semantic attribute that are correctly classified.

Table 6.9: Impact of different batch sizes on certain attributes – The positive recall for selected attributes and varying batch sizes are presented. The reduction of the training batch size is beneficial for rare attributes with low intra-class diversity in training data, such as *LowerStripe*. However, the recognition of attributes with large variations in appearance like *Hat* and equally balanced attributes such as the gender deteriorates for small batches.

B	Market-1501			PA-100K		
	downblue	*downgray*	*gender*	*LowerStripe*	*Hat*	*Femal*
64	56.1	58.0	91.3	60.0	**49.5**	91.0
32	56.2	58.9	**91.6**	60.0	46.7	**91.1**
16	**60.1**	**59.8**	91.1	**64.0**	42.9	89.7

On the small Market-1501 dataset, reducing the batch size improves the recognition of positive samples of soft biometrics such as the lower-body colors blue (*downblue*) and gray (*downgray*) with limited intra-class diversity and fuzzy transitions to other colors. For instance, using batches of size 16 instead of 64 improves the positive recall of the attribute *downblue* by 4 percentage points. In contrast, equally balanced attributes such as gender suffer from the reduced batch size as the model overfits on nuances instead of learning the overall concept behind that attribute. However, on average, more attributes benefit and the overall results improve.

Concerning PA-100K, there are also very unbalanced attributes with few examples, *e.g.*, *LowerStripes*, which refers to a person wearing striped lower-body clothing. This attribute appears in only 0.5% of training and 0.2% of test images, respectively. As a result, the positive recall increases for smaller batch sizes, since otherwise the few positive samples are outweighed by negative ones. However, analogous to the results on the Market-1501 dataset, the recognition of the gender attribute *Femal* deteriorates. In addition, soft biometric characteristics that appear in a wide variety of types and appearances, such as *Hat*, which covers all kinds of headwear, suffer and positive recall drops sharply. The model fails to learn the overall concept behind the attribute due to the strong influence on batch updates from single instances of headwear. The loss for the attribute fluctuates during training and hardly converges.

Next, the influence of batch size on generalization is investigated using the UPAR dataset. In general, the literature shows stronger generalization of models trained with smaller batches [Kes16, Hof17]. The results of the experiments are given in Table 6.10.

Table 6.10: Generalization results for varying batch sizes – Analogous to the specialization results, training with large amounts of training data (LOOCV protocol) benefits from larger batch sizes than training with less data (4FCV protocol). While the best performance concerning the LOOCV evaluation protocol is achieved by training the PAR model with batches of size 64, applying different batch sizes for the smaller and larger splits is optimal for the 4FCV protocol. This is referred to as 64/16.

B	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
64	**74.2**±2.4	**83.2**±2.1	**60.1**±5.1	**20.0**±5.3	**22.0**±9.1
32	73.7±2.6	83.0±2.1	59.2±5.0	18.9±4.9	20.3±8.9
16	73.0±2.5	82.1±2.3	56.8±5.0	17.1±4.5	18.3±8.4
8	71.1±2.7	81.1±2.6	53.4±5.1	14.4±3.7	16.1±8.2
B	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
64	69.5±2.0	78.3±3.7	49.3±6.6	12.4±4.4	13.7±5.1
32	69.3±1.7	78.2±3.8	49.3±6.4	12.2±4.1	13.4±4.8
16	69.1±1.3	78.0±3.8	49.0±5.8	11.7±3.6	13.4±4.8
8	68.2±1.4	77.3±4.1	47.2±5.1	10.5±2.9	11.8±3.7
64/16	**69.9**±1.6	**78.4**±3.8	**50.1**±6.1	**12.7**±4.3	**14.1**±5.0

The observations are consistent with those obtained for the specialization case on the individual datasets. When enough training data and, hence, diversity is available, which is the case for the LOOCV evaluation protocol, choosing a batch size of 64 is most appropriate to achieve the best outcomes for the PAR and attribute-based retrieval tasks. Indeed, this batch size leads to the strongest performance on each of the cross-validation splits without exception.

When only a single dataset is leveraged for training in the 4FCV protocol, *i.e.*, less training data is available, the decrease in accuracy due to reduced batch sizes is smaller. However, the best results are obtained when the findings from

the specialization datasets are transferred and different batch sizes of 64 and 16 are chosen for the two large and small splits, respectively.

The results on the optimal batch size are consistent across the specialization and generalization cases. Thus, the hypothesis that an increased number of batch updates through a decreased batch size is beneficial for generalization is supported [Hof17].

6.1.6 Dropout

Training deep neural networks on datasets of limited size and diversity carries the risk of also learning the statistical noise contained in the training dataset and, thus, overfitting on that data, rather than learning deep features that generalize well to novel data. This problem becomes more severe with increasing model size and decreasing number of training images. Srivastava et al. [Sri14] propose dropout as a regularization technique since they found that neurons can co-adapt to correct and compensate for the failures of other nodes. Since these complex co-adaptations are data specific, generalization capability suffers.

Dropout randomly ignores a specified ratio of connections r_{drop} between two layers, simulating different architectures during training. This makes the training process noisy and avoids co-adaptation due to different connections between the layers in each forward pass. Instead, network nodes must take on more or less responsibility for the output depending on the connections that are dropped. This reduces the risk of overfitting the training images and increases the robustness of the predictions, as the model learns to produce meaningful outputs even if some information is missing or unreliable.

For the experiments, a dropout layer is positioned between the GAP and the FC classification layer, *i.e.*, inputs from the backbone to the classifier are dropped probabilistically. Different dropout rates r_{drop} are explored based on the best approach from the previous section.

Experimental results for the specialization datasets are provided in Table 6.11.

Table 6.11: Impact of dropout on the specialization results – The use of dropout improves PAR as well as attribute-based person retrieval. The optimal dropout rate r_{drop} depends on the dataset.

r_{drop}	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
0.0	86.8	88.5	**66.3**	26.9	26.1	84.3	89.9	73.4	30.4	**38.9**
0.1	**86.9**	**88.6**	66.2	**27.4**	**26.5**	84.4	89.9	73.4	30.5	38.4
0.2	86.6	88.4	65.5	26.9	26.2	84.7	89.9	73.5	30.6	38.4
0.3	86.7	88.5	65.4	27.2	26.7	**84.8**	**90.0**	**73.6**	30.6	37.3
0.4	86.4	88.2	65.0	26.5	26.2	84.7	89.9	**73.6**	**30.7**	38.3
0.5	86.6	88.2	64.5	26.0	25.6	84.6	89.9	73.4	30.6	38.3
r_{drop}	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
0.0	**79.2**	80.9	**61.6**	22.3	15.9	80.9	86.6	70.9	37.0	50.0
0.1	**79.2**	**81.0**	61.3	**22.6**	**16.3**	81.0	86.7	71.2	37.6	50.8
0.2	79.0	**81.0**	61.2	22.5	16.1	81.1	86.8	71.6	38.0	51.9
0.3	78.6	**81.0**	61.1	22.5	**16.3**	**81.4**	87.0	71.9	38.8	52.9
0.4	78.5	**81.0**	61.1	22.3	15.8	**81.4**	87.0	72.2	39.3	**53.7**
0.5	77.9	80.9	60.9	22.3	15.9	80.9	**87.1**	**72.4**	**39.5**	52.3

In general, introducing dropout as regularization technique improves the results. The experiments suggest that the optimal choice of the dropout rate depends on the data. While $r_{\text{drop}} = 0.1$ gives the best results for the PETA and the RAPv2 dataset, higher values are beneficial for PA-100K and Market-1501. The latter two seem to contain more statistical noise and, therefore, an increased tendency of models to overfit on these datasets.

Based on these observations, the datasets are divided into two classes with different dropout rates. For models trained with the PETA and RAPv2 datasets, 0.1 is set as the optimal dropout configuration, while the other two datasets use a value of 0.4, as they seem to have an increased risk of overfitting and co-adaptation of nodes.

Moreover, reducing overfitting to the training data is expected to improve the generalization capabilities on another domain. Looking at the UPAR results in Table 6.12, the findings support this. All metrics exhibit small improvements through the use of dropout.

Table 6.12: Impact of dropout on the generalization results – Utilizing dropout is beneficial to avoid overfitting and improve the PAR and attribute-based person retrieval results in terms of generalization. The 4FCV evaluation protocol that uses fewer data for training profits from higher dropout rates r_{drop} than the LOOCV protocol.

r_{drop}	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
0.0	74.2±2.4	83.2±2.1	60.1±5.1	20.0±5.3	22.0±9.1
0.1	74.3±2.3	83.2±2.0	60.2±5.2	20.2±5.2	22.0±9.2
0.2	**74.4**±2.3	**83.4**±2.0	**60.4**±5.1	**20.3**±5.2	**22.2**±9.0
0.3	74.3±2.2	83.3±2.0	60.3±5.0	**20.3**±5.3	22.0±9.1
0.4	74.1±2.3	83.3±2.0	60.2±5.0	20.2±5.3	22.1±9.0
0.5	74.0±2.3	83.3±2.0	60.1±5.1	20.0±5.3	21.9±9.2
r_{drop}	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
0.0	69.9±1.6	78.4±3.8	50.1±6.1	12.7±4.3	14.1±5.0
0.1	**70.1**±1.6	78.7±3.7	50.4±6.2	**12.9**±4.3	**14.3**±4.8
0.2	70.0±1.7	78.7±3.7	50.5±6.1	**12.9**±4.3	**14.3**±5.1
0.3	**70.1**±1.7	78.7±3.7	50.5±6.1	**12.9**±4.3	**14.3**±5.1
0.4	**70.1**±1.8	78.7±3.8	**50.6**±6.0	**12.9**±4.3	**14.3**±5.1
0.5	70.0±1.7	**78.8**±3.7	50.5±6.1	12.8±4.2	14.2±5.0

Regarding the optimal dropout rate, the 4FCV evaluation protocol, which has only one data source for training and, thus, a higher risk of overfitting, benefits from higher dropout rates compared to the LOOCV scheme, which leverages data from multiple sources during training. Therefore, dropout rates of 0.2 and 0.4 are chosen for the LOOCV and the 4FCV procedure, respectively.

6.1.7 Optimizer

Numerous research utilizes adaptive optimizers such as Adam [Kin14] for training CNNs due to their reduced need for costly hyperparameter tuning. Moreover, such optimizers typically result in faster convergence and, thus, shorter training time than, *e.g.*, stochastic gradient descent. However, it has been found that models trained with Adam tend to have suboptimal generalization capabilities.

Loshchilov et al. [Los17] highlight one potential explanation for this phenomenon, attributing it to the reduced effectiveness of L2 weight regularization in implementations of Adam. To address this issue, they introduced the AdamW optimizer, which is specifically designed to rectify this drawback by fixing the weight regularization. Despite this advancement, prevailing PAR methods [Jia21b] continue to rely on the standard version of Adam and consequently suffer from poor generalization.

To close this gap, the use of AdamW in the context of PAR is investigated, particularly in the case of generalization. Further algorithms, *e.g.*, RAdam [Liu19], were also examined but achieved worse results than AdamW. Therefore, only results for AdamW are reported. The results were obtained by building on the best approaches from the previous Section 6.1.6.

The results obtained for the specialization datasets in Table 6.13 demonstrate clear improvements when utilizing the AdamW optimizer rather than the standard Adam optimizer. This improvement is consistent across both tasks and is reflected in all metrics evaluated, with a few isolated exceptions. However, it is noteworthy that using AdamW only produces positive effects when combined with SWA and adjusted batch size for small datasets. Otherwise, it fails to generate consistent improvement in comparison with Adam. This may imply that the enhanced weight regularization in AdamW alone is unable to entirely compensate for overfitting of weights to the training data. In contrast, AdamW appears to further improve performance of a well-regularized and configured model.

Table 6.13: Comparison of optimizers for the case of specialization – The results clearly indicate the superiority of the AdamW optimizer over the Adam optimizer. In particular, the attribute-based person retrieval task benefit from the use of this optimizer.

Optimizer	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Adam	86.9	88.6	**66.2**	27.4	**26.5**	84.7	89.9	73.6	30.7	38.3
AdamW	**87.3**	**89.0**	**66.2**	**28.1**	26.4	**84.8**	**90.1**	**74.1**	**31.5**	**40.1**
Optimizer	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Adam	79.2	81.0	61.3	22.6	16.3	**81.4**	**87.0**	72.2	39.3	53.7
AdamW	**80.9**	**81.1**	**61.7**	**23.3**	**16.9**	81.0	**87.0**	**72.6**	**39.6**	**54.6**

The observations made regarding the effectiveness of the AdamW optimizer remain identical when applied to the UPAR dataset. The specific results showing the impact of the optimizer can be found in Table 6.14.

Table 6.14: Comparison of optimizers for the case of generalization – The use of the AdamW optimizer is advantageous for both generalization protocols and tasks.

Optimizer	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
Adam	74.4±2.3	83.4±2.0	60.4±5.1	20.3±5.2	22.2±9.0
AdamW	**75.0**±2.0	**83.8**±2.0	**61.3**±5.4	**21.3**±5.8	**23.0**±9.7
Optimizer	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
Adam	70.1±1.8	78.7±3.8	50.6±6.0	12.9±4.3	14.3±5.1
AdamW	**70.5**±1.9	**79.2**±3.7	**51.2**±6.3	**13.5**±4.5	**15.0**±5.2

The use of AdamW consistently improves the results for both cross-domain generalization protocols. Thus, it can be concluded that the substitution of Adam with AdamW has an apparent positive impact, underscoring its potential to significantly improve model performance and generalization capabilities, especially in cross-domain scenarios.

6.1.8 Label Smoothing

Label smoothing, as proposed by Szegedy et al. [Sze16], is a regularization technique which is designed to improve generalization, increase the robustness, and mitigate overconfidence of neural networks. Its application involves the modification of training labels during the training process.

In its original form, label smoothing is applied to multi-class classification tasks where training labels are typically one-hot encoded. This means that a single element of the label vector is 1 to indicate the correct class, while the rest are 0. The basic concept of label smoothing is to introduce a controlled amount of uniform noise into these labels. Instead of using 1 as the target label

of the true class and 0 for the others, label smoothing distributes a fraction of noise among all classes.

However, PAR is a multi-label classification problem, *i.e.*, multiple classes per image can be true simultaneously. Therefore, an adaptation of the concept is required. The adapted version of label smoothing for multi-label attribute recognition works on a per-attribute basis, adjusting the confidence of each attribute's target value as follows:

$$y_{ij}^{\mathrm{LS}} = (1 - \alpha_{\mathrm{LS}})y_{ij} + \alpha_{\mathrm{LS}}(1 - y_{ij}). \tag{6.4}$$

The smoothed target value y_{ij}^{LS} for the j-th attribute in the i-th image is calculated using the parameter α_{LS}. This parameter introduces a controlled degree of uncertainty into the annotated target values y_{ij}, effectively tempering the certainty of the model. For instance, if the parameter α_{LS} is set to 0.1, the target values for positive samples are adopted to 0.9 and for negative samples to 0.1.

By adopting label smoothing, the model is discouraged from becoming overconfident and overly reliant on the training labels. Instead, it is encouraged to capture a more generalized and smoother decision boundary, thus minimizing the risk of overfitting. Given the inherent noise and occasional incorrect labels present in PAR training data due to the strenuous and often subjective annotation processes, the use of label smoothing is considered a potentially beneficial strategy. Furthermore, mitigating overconfidence is expected to have a positive impact on attribute-based person retrieval, since the classifier's output scores play a crucial role in the ranking process and should, therefore, provide reliable estimates of the true presence probabilities of attributes in images.

Table 6.15 explores the effect of different values of the smoothing parameter α_{LS}. The analysis provides insight into the behavior of PAR models under different degrees of label smoothing. The results were obtained using the models trained with AdamW as basis, as described in the preceding section.

Table 6.15: Impact of label smoothing on the specialization results – Using label smoothing improves the results on each of the datasets. However, differing values for the degree of label smoothing α_{LS} lead to optimal performance concerning the PAR and retrieval task, respectively. Datasets such as PETA and Market-1501, which contain instance-wise annotations and, thus, increased noise in the labels, benefit from stronger label smoothing.

α_{LS}	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
0.00	87.3	89.0	66.2	28.1	26.4	84.8	90.1	74.1	31.5	40.1
0.05	87.4	89.1	66.6	28.3	26.3	84.6	90.1	74.0	31.6	40.6
0.10	87.2	89.2	67.4	29.1	27.4	84.8	90.1	73.6	**31.7**	**41.3**
0.15	87.3	89.4	68.1	**29.4**	27.7	84.8	90.2	74.1	31.4	40.2
0.20	87.6	89.5	**68.4**	29.3	**28.3**	85.0	**90.3**	**74.2**	31.1	39.5
0.25	**88.0**	**89.6**	68.1	28.3	27.1	**85.2**	90.2	73.8	30.0	37.7
α_{LS}	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
0.00	**80.9**	81.1	61.7	23.3	16.9	81.0	87.0	72.6	39.6	54.6
0.05	80.8	81.1	**61.9**	**23.6**	**17.0**	81.1	87.1	72.6	39.6	53.5
0.10	80.7	81.1	61.8	23.3	16.9	81.6	87.4	72.7	40.1	54.3
0.15	80.5	81.2	61.6	23.0	16.7	81.7	87.4	**73.2**	**40.8**	**56.0**
0.20	80.0	**81.3**	61.2	22.2	16.0	81.9	87.6	73.1	40.5	54.1
0.25	80.0	**81.3**	60.5	21.2	15.2	**82.0**	**87.8**	72.9	39.9	52.9

The evaluation confirms that the application of label smoothing generally improves the results. However, the extent of this improvement varies from dataset to dataset, highlighting a dependency on the intrinsic properties of the specific dataset. Furthermore, the results indicate a discrepancy between optimal parameter choices for PAR and person retrieval. Observations show that a higher degree of smoothing improves PAR results. Both label-based mA and instance-based F1 improve for increasing values of α_{LS}, except for the mA on the RAPv2 dataset. This indicates a substantial amount of annotation errors in the datasets and a strong tendency of the models to overfit the training data. In contrast, retrieval metrics begin to drop again for high degrees of uncertainty added to the target labels during training. Excessive uncertainty beyond the degree actually present in the training data seems to

result in underconfident models. As a result, the classifier's output probabilities are skewed, computed retrieval distances are less reliable, and, therefore, retrieval accuracy decreases.

In addition, the application of label smoothing yields notable and consistent improvements in the context of R-1 accuracy. Typically, the R-1 metric exhibits fluctuations and, sometimes behaves inconsistently with the other retrieval metrics due to its emphasis on only the top-ranked matching gallery image. Observations demonstrate that label smoothing has a stabilizing effect on this metric, consistently improving its performance. Overconfidence poses a significant problem for this metric, as even subtle variations in a model's prediction scores can lead to swapped positions in the final retrieval ranking. This phenomenon underscores the importance of mitigating overconfidence through techniques such as label smoothing or calibration (see Section 7.2.1), to ensure the robustness and reliability of attribute-based person retrieval.

The experiments show that models trained on the PETA and Market-1501 datasets require a higher degree of label smoothing than the other datasets to achieve optimal performance. This is due to the instance-wise annotations. There is no guarantee that the individual image actually matches with the attribute labels provided. As a result, the certainty of the labels is lower than for datasets where each images is annotated individually. Based on these observations, it was decided to set α_{LS} to 0.15 for the datasets with instance-wise annotations and 0.05 for those containing image-wise labels.

Similar observations are made for the UPAR dataset, as shown in Table 6.16. In the context of the LOOCV evaluation protocol, a lower value of 0.15 for the smoothing parameter appears to be optimal for achieving peak performance. On the other hand, when using the 4FCV protocol, which involves training with less data, a slightly higher level of regularization is required, and 0.2 proves to be the optimal choice. Increasing the level of smoothing beyond the reported values in the table does not provide added advantages in terms of retrieval.

Table 6.16: Impact of label smoothing on the generalization results – The use of label smoothing improves the generalization performance concerning attribute recognition as well as retrieval. Besides an increase in performance, reduced standard deviation across the splits is observed, especially for the 4FCV evaluation protocol.

α_{LS}	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
0.00	75.0±2.0	83.8±2.0	**61.3**±5.4	21.3±5.8	23.0±9.7
0.05	74.9±2.0	**83.9**±2.0	**61.3**±5.3	21.3±5.9	23.1±9.8
0.10	74.9±2.2	**83.9**±1.9	**61.3**±5.3	21.5±5.8	23.1±9.3
0.15	75.1±2.0	**83.9**±1.9	61.2±5.3	**21.6**±5.8	**23.4**±9.4
0.20	**75.2**±2.0	**83.9**±1.9	61.0±5.3	21.2±5.7	23.2±9.5
α_{LS}	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
0.00	70.1±2.3	79.2±3.7	50.7±6.5	13.5±4.5	15.2±5.3
0.05	70.3±1.9	79.3±3.6	50.9±6.3	13.6±4.4	15.3±5.1
0.10	70.6±1.6	79.5±3.5	51.2±6.2	**13.8**±4.3	**15.5**±4.9
0.15	70.7±1.6	79.6±3.4	**51.7**±6.0	**13.8**±4.3	15.3±5.0
0.20	**70.9**±1.6	**79.7**±3.4	**51.7**±5.8	**13.8**±4.1	15.4±4.9

Moreover, utilizing this regularization technique reduces fluctuation across the splits, particularly evident in the more demanding 4FCV protocol. Comparing the results achieved with and without label smoothing, it is clear that in addition to increasing metrics, standard deviations observed among the splits are reduced. Challenging splits with lower evaluation scores benefit greatly by label smoothing, resulting in improved performance that catches up to that of the easier splits. Therefore, it is concluded that incorporating label smoothing is crucial in challenging scenarios where only limited training data with significant label noise is available.

6.1.9 Summary

Several optimization concerning the PAR baseline were investigated for their impact on attribute-based person retrieval. Initially, treating semantic attributes as binary is justified by showcasing superior attribute-based person

retrieval performance, compared to considering multi-class attributes explicitly. Various backbones were compared afterward. The results indicate that contemporary CNNs, specifically ConvNeXt-Base [Liu22], offer a suitable balance between retrieval accuracy and inference time. While the Swin-Large [Liu21] transformer outperforms the ConvNeXt architecture for the specialization case, ConvNeXt-Base is faster and achieves comparable or even better accuracy in generalization. Thus, ConvNeXt-Base is selected as the default backbone for the further experiments in this thesis. The examination of SWA techniques suggests more robust and smoother training behavior and notable improvements in both specialization and generalization performance. Particularly, the parameter-free SMA [Arp22] proved to be advantageous for both PAR and retrieval. A comparison between the cross-entropy and the focal loss [Lin17] functions reveals minimal influence in PAR, but the focal loss function clearly enhances the attribute-based person retrieval quality. This suggests better model calibration, indicating the advantages of the focal loss function. Reducing the batch size for small datasets increases the number of batch updates for the models, which leads to less overfitting and improved recognition of rare attributes, since the impact of such attributes' samples is enlarged. Applying dropout [Sri14] also leads to slight improvements by avoiding the co-adaption of model weights and, consequently, overfitting. The analysis of the Adam [Kin14] and AdamW [Los17] optimizers demonstrates that both tasks benefit from AdamW. Notably, this applies only if AdamW is utilized after regularizing the model, *i.e.*, it is used in conjunction with SMA, optimal batch sizes, and dropout. This finding suggests that an appropriately regularized model training is necessary for AdamW to fully demonstrate its strengths. Additionally, the study explores the use of label smoothing [Sze16] to prevent over- and underconfidence of PAR models, resulting in improved outcomes for both tasks and evaluation cases. Datasets with instance-level annotations experience benefits from introducing greater uncertainty to the training labels for loss calculation, as this reflects the enlarged uncertainty inherent to annotations for the individual images. Moreover, increasing the hyperparameter α_{FL} leads to further enhancement in PAR. However, the performance of attribute-based person retrieval declines when it comes to high values. This is due to models switching between

over- and underconfident states, when the level of introduced uncertainty by label smoothing surpasses the true uncertainty of ground truth annotations. As a result, the calibration of the attributes' confidence scores is impaired and, consequently, retrieval is adversely affected.

Table 6.17 summarizes the results achieved through the optimizations for the specialization datasets. In terms of mADM, particularly SMA and the use of the focal loss function lead to consistent improvements across the datasets. Furthermore, reducing the batch size and implementing label smoothing proved to be significant for small datasets with instance-wise annotations such as the PETA and Market-1501 datasets.

Table 6.17: Summary of PAR model optimization – Results for the specialization case are provided. The baseline refers to the approach described in Chapter 5 with ConvNeXt-Base as the backbone model. Especially, SMA and the use of the focal loss function yield consistent improvements. Furthermore, small datasets with instance-level annotations benefit notably from adjusting the batch size and employing label smoothing.

Approach	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	86.1	88.1	61.8	24.4	24.4	82.2	88.5	68.8	26.2	34.5
+ SMA	86.4	88.4	63.4	25.5	25.1	84.3	90.0	72.3	30.4	38.8
+ Focal loss	86.4	88.4	65.1	27.0	26.3	84.3	89.9	73.4	30.4	38.9
+ Batch size	86.8	88.5	66.3	26.9	26.1	84.3	89.9	73.4	30.4	38.9
+ Dropout	86.9	88.6	66.2	27.4	26.5	84.7	89.9	73.6	30.7	38.3
+ AdamW	**87.3**	89.0	66.2	28.1	26.4	**84.8**	**90.1**	**74.1**	31.5	40.1
+ Label smooth.	**87.3**	**89.4**	**68.1**	**29.4**	**27.7**	84.6	**90.1**	74.0	**31.6**	**40.6**
Approach	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	79.3	80.0	58.5	20.5	14.5	80.7	85.7	65.9	31.6	47.7
+ SMA	79.4	**81.1**	60.6	22.6	16.3	79.7	86.6	67.9	34.0	50.6
+ Focal loss	79.2	80.9	61.1	22.3	15.9	80.1	86.4	69.7	35.5	50.2
+ Batch size	79.1	80.9	61.1	22.3	15.9	80.9	86.6	70.9	37.0	50.0
+ Dropout	79.2	81.0	61.3	22.6	16.3	81.4	87.0	72.2	39.3	53.7
+ AdamW	**80.9**	**81.1**	61.7	23.3	16.9	81.0	87.0	72.6	39.6	54.6
+ Label smooth.	80.8	**81.1**	**61.9**	**23.6**	**17.0**	**81.7**	**87.4**	**73.2**	**40.8**	**56.0**

Similar outcomes are noted for both evaluation protocols of the UPAR dataset, as presented in Table 6.18.

Table 6.18: Summary of PAR model optimization for the UPAR dataset – Results for the generalization case are provided. The baseline refers to the approach described in Chapter 5 with ConvNeXt-Base as the backbone model. Similar to Table 6.17, SMA, the focal loss function, and AdamW lead to the most prominent improvements.

Approach	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
Baseline	73.2±1.9	82.1±2.7	56.1±5.1	17.6±4.3	19.2±8.3
+ SMA	74.0±2.4	83.4±2.0	58.5±5.4	19.5±5.4	21.4±9.3
+ Focal loss	74.2±2.4	83.2±2.1	60.1±5.1	20.0±5.3	22.0±9.1
+ Batch size	74.2±2.4	83.2±2.1	60.1±5.1	20.0±5.3	22.0±9.1
+ Dropout	74.4±2.3	83.4±2.0	60.4±5.1	20.3±5.2	22.2±9.0
+ AdamW	75.0±2.0	83.8±2.0	**61.3**±5.4	21.3±5.8	23.0±9.7
+ Label smoothing	**75.1**±2.0	**83.9**±1.9	61.2±5.3	**21.6**±5.8	**23.4**±9.4
Approach	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
Baseline	70.1±1.5	77.2±4.2	46.4±5.8	11.0±3.6	12.8±4.6
+ SMA	69.5±1.9	78.4±3.8	47.8±6.4	11.8±4.3	13.2±5.3
+ Focal loss	69.5±2.0	78.3±3.7	49.3±6.6	12.4±4.4	13.7±5.1
+ Batch size	69.9±1.6	78.4±3.8	50.1±6.1	12.7±4.3	14.1±5.0
+ Dropout	70.1±1.8	78.7±3.8	50.6±6.0	12.9±4.3	14.3±5.1
+ AdamW	70.5±1.9	79.2±3.7	51.2±6.3	13.5±4.5	15.0±5.2
+ Label smoothing	**70.9**±1.6	**79.7**±3.4	**51.7**±5.8	**13.8**±4.1	**15.4**±4.9

In conclusion, the findings demonstrate that global image-based PAR methods possess tremendous potential. Additionally, the systematic investigation of design choices that includes recent advancements in deep learning exhibits remarkable improvement compared to the baseline approach. Consequently, a new strong baseline is obtained, which is suitable for specialization as well as generalization.

6.2 Normalization

Underrepresented attributes in training data present a considerable challenge, which is commonly tackled through weighted loss functions. Nonetheless,

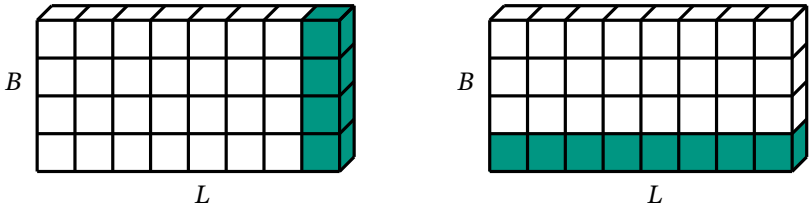
the effectiveness of this approach depends on the proper parameterization. Usually, a model that is trained without a positive ratio-based loss function underestimates the number of attributes present in an image. This is due to the uncertainty associated with attributes that have only a few training samples. Limited diversity hinders the model from learning the abstract concepts underlying these attributes. Consequently, the decision threshold is often not surpassed for positive samples in the test set. This issue particularly arises in generalization tasks, where large differences exist between the training and test data, and varying underlying distributions of the attributes are present in these sets. Although emphasizing a few positive samples during training enhances the recall of positive samples, it incurs additional false positives. The model's overconfidence leads to an overestimation of the number of attributes present in the input image. Consequently, while the mA improves with correct recognition of more positive samples, the instance-based precision is negatively impacted due to a rise in false positive predictions, causing a decrease in instance-based F1. Furthermore, overconfident models underperform on the retrieval task since their outputs, used to calculate retrieval distances, do not align with empirical probabilities for the presence of the attributes.

Research in the field indicates that algorithms typically optimize either label-based metrics or instance-based metrics. Different network structures and added components, such as attention mechanisms, can impact these metrics differently. These findings align with research by the author [Cor23] and are consistent with the results reported in this thesis, which demonstrate a similar discrepancy in performance metrics. For example, when using ConvNeXt-Base as the backbone, significantly stronger results concerning the instance-based F1 are achieved than for the label-based mA on the Market-1501 dataset. In general, instance-based measures are better predictors of attribute-based person retrieval performance, and models with strong performance in this type of PAR metric outperform those optimizing label-based scores [Cor23]. However, achieving promising results regarding both types of metrics and retrieval is optimal. This study assumes that balancing label- and instance-based metrics in PAR has great potential in enhancing the quality of retrieval. To achieve this, a novel normalization strategy for PAR models is proposed by the author of this thesis [Spe23a].

6.2.1 Methodology

To alleviate the problems caused by imbalanced attribute distributions without harming attribute-based person retrieval performance and to obtain balanced PAR results, the PARNorm module is proposed. The core idea is to sequentially normalize the classifier's output logits $\mathbf{x}_i$ concerning the attributes and afterward regarding the entire person description. The PARNorm module expands the work of Zhao et al. [Zha18b], which only applies attribute-wise normalization.

Both types of normalization operations are visualized in Figure 6.5 and resemble batch [Iof15] and layer normalization [Ba16], respectively. Each subfigure visualizes a training batch consisting of B logit vectors $\mathbf{x}_i$ of dimension L, which corresponds to the number of attributes to recognize. The elements highlighted in green depict the elements that are considered for normalization.



(a) Attribute-wise **(b)** Instance-wise

Figure 6.5: PARNorm – Each figure displays a batch of attribute logit vectors. Initially, attribute-wise normalization is conducted to enhance the recognition of imbalanced attributes, as illustrated in Figure 6.5a. Moreover, it is suggested to subsequently normalize the output logits along the attribute dimension, as shown in Figure 6.5b. This aims at improving instance-based metrics by avoiding the over- or underestimation of the number of a person's semantic attributes.

First, logits are normalized in an attribute-wise manner, as shown in Figure 6.5a. In practice, the classifier is often not confident about positive samples of attributes that appear in only a few training images. As a consequence, the predicted output probabilities $\mathbf{p}_i$ for positive samples on unseen data tend

to be low and may not surpass the decision threshold. In addition, the output distributions for both positive and negative samples tend to be close together, leading to difficulties in discrimination between the manifestations of the attribute. Normalizing the logits per attribute across the batch samples enhances the discrimination between negative and positive samples. This process widens the range of logit values and, therefore, improves the recognition of the attribute. It involves calculating the mean μ_j^B and corresponding variance $\sigma_j^{2,B}$ of logits values per attribute within a batch as follows:

$$\mu_j^B = \frac{1}{B}\sum_{i=1}^{B} x_{ij} \tag{6.5}$$

and

$$\sigma_j^{2,B} = \frac{1}{B}\sum_{i=1}^{B}(x_{ij} - \mu_j^B)^2. \tag{6.6}$$

Here, B represents the mini-batch size. Based on these statistics, the batch-normalized logits $\hat{x}_{i,j}^B$ for the j-th attribute of the i-th sample in the batch are then computed as

$$\hat{x}_{ij}^B = \gamma_j^B \left(\frac{x_{ij} - \mu_j^B}{\sqrt{\sigma_j^{2,B} + \epsilon}} \right) + \beta_j^B, \tag{6.7}$$

where ϵ is a small constant introduced for numerical stability, and γ_j^B and β_j^B are learnable scale and shift parameters, respectively. These parameters control the magnitude of the outputs.

However, attribute-wise normalization calibrates the logits in such a way that the positive recall of rarely present attributes increases with the risk of also introducing false positive predictions. Thus, particularly the label-based mA benefits from this kind of normalization. On the other hand, an overestimation of the number of attributes for a person instance may impair instance-based metrics and lead to incorrect retrieval results. Therefore, instance-wise

normalization is applied afterward to compensate for this effect. As can be seen in Figure 6.5b, instance-wise normalization is performed along the attribute dimension, *i.e.*, separately for each instance included in the batch. The formulation is similar to the first normalization technique, but the activations are normalized across the attributes in a per instance manner as

$$\hat{x}_{ij}^{L} = \gamma_i^L \left(\frac{x_{ij} - \mu_i^L}{\sqrt{\sigma_i^{2,L} + \epsilon}} \right) + \beta_i^L \tag{6.8}$$

with

$$\mu_i^L = \frac{1}{L} \sum_{j=1}^{L} x_{ij} \tag{6.9}$$

and

$$\sigma_i^{2,L} = \frac{1}{L} \sum_{j=1}^{L} (x_{ij} - \mu_i^L)^2 \tag{6.10}$$

being the mean and variance across the attribute logits for the i-th individual in the batch.

6.2.2 Evaluation

The proposed PARNorm module is evaluated using ConvNeXt-Base as backbone. The baseline approach without modification is leveraged for comparison.

First, an ablation study is conducted to examine the influence of the two normalization techniques on PAR and retrieval results. Quantitative results for the PETA dataset are provided in Table 6.19. As expected, solely applying attribute-wise normalization mainly leads to an increase in mA. The reason for this is that the presence of attributes is identified correctly by the model to a much higher portion. The positive recall across all attributes is only

77.6% for the baseline approach without normalization, whereas including attribute-wise normalization improves this score to 82.2%. However, at the same time, additional false positive recognitions are produced, which deteriorate the instance-based precision by 1.5 points. Thus, the instance-based F1 score is slightly worse compared to the baseline. Regarding the retrieval task, only the mADM measure improves through attribute-wise normalization, while the other evaluation metrics show worse performance. This finding mostly transfer to the remaining datasets. Deteriorated mAP and R-1 scores are also observed for those datasets. For mADM, however, findings are not consistent across the datasets. On the PETA and PA-100K datasets, the mADM improves, while attribute-wise normalization leads to degraded performance for the Market-1501 and RAPv2 datasets. Since a more severe drop in instance-based F1 is observed, this indicates that the number of additional false positives is much larger for the latter datasets. As a result, this also negatively impacts the retrieval performance measured by mADM. The mADM measure considers the ratio of matching attributes between the query and gallery images, which is why the increased positive recall may improve this metric even if there are some additional false positives. Images depicting challenging attributes receive lower distances for those attributes due to better recognition and, therefore, are ranked in earlier positions. However, if the presence of attributes is overestimated to a larger extent, as on RAPv2 and Market-1501, this also applies to incorrect gallery samples showing different sets of attributes.

Table 6.19: PARNorm ablation study – Baseline results, results for the single normalization techniques, and for the combination of both are presented for the PETA dataset. Performing attribute-wise normalization followed by instance-wise normalization, as proposed in this thesis, leads to the best performance concerning each of the metrics.

Normalization		PETA				
Attribute-wise	Instance-wise	mA	F1	mADM	mAP	R-1
		86.1	88.1	61.8	24.4	24.4
✓		88.3	87.8	62.8	23.7	23.8
	✓	87.8	88.6	64.1	26.1	26.3
✓	✓	**88.6**	**88.7**	**65.3**	**26.8**	**26.5**

The application of instance-wise normalization improves all metrics and is especially beneficial for retrieval. Normalizing the scores per instance balances the confidence scores of the classifier across the attributes, which improves the computation of the retrieval distances. The impact of single attributes on the resulting distance is much more balanced as without this normalization.

Combining both normalization components sequentially, melds the benefits and achieves the best results. Label-based mA and instance-based F1 are balanced and, particularly, the mA is enhanced. As assumed, the improved balance also leads to enhanced retrieval performance. Measured by mADM, the quantitative improvement is 3.5 points for the PETA dataset.

These findings transfer to the other datasets, as is shown in Table 6.20. Attribute-based person retrieval showcases improvements ranging from 1.2 to 3.5 points in mADM for the datasets. Moreover, the normalization module provides benefits to both types of PAR metrics.

Table 6.20: Specialization results for the PARNorm module – The use of the PARNorm module consistently improves the PAR and person retrieval results for all datasets. Furthermore, the results indicate that the normalization module is able to enhance the mA metric greatly without negative effects concerning instance-based F1.

Normalization	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	86.1	88.1	61.8	24.4	24.4	82.2	88.5	68.8	26.2	**34.5**
PARNorm	**88.6**	**88.7**	**65.3**	**26.8**	**26.5**	**85.6**	**89.1**	**71.3**	**28.9**	34.2
Normalization	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	79.3	80.0	58.5	20.5	14.5	80.7	85.7	65.9	31.6	**47.7**
PARNorm	**81.2**	**80.3**	**59.7**	**22.0**	**16.0**	**84.7**	**86.1**	**68.3**	**34.7**	47.5

As reported in Table 6.21, the results for the UPAR dataset indicate that using the PARNorm module leads to a decline in instance-based F1 score in the generalization case. Nevertheless, remarkable improvements are observed in terms of mA. Moreover, the retrieval metrics increase for both evaluation protocols, highlighting the benefits of the PARNorm module for this task.

Table 6.21: Generalization results for the PARNorm module – Similar to the specialization results, utilizing the PARNorm module leads to an improvement in mA. However, this is accompanied by a decrease in the instance-based F1. Nevertheless, the attribute-based retrieval performance is enhanced.

Normalization	UPAR LOOCV				
	mA	F1	mADM	mAP	R-1
Baseline	73.2±1.9	**82.1**±2.7	56.1±5.1	17.6±4.3	19.2±8.3
PARNorm	**76.8**±1.8	81.6±2.4	**57.4**±4.8	**18.6**±4.7	**20.2**±8.8
Normalization	UPAR 4FCV				
	mA	F1	mADM	mAP	R-1
Baseline	70.1±1.5	**77.2**±4.2	46.4±5.8	11.0±3.6	12.8±4.6
PARNorm	**73.0**±1.6	76.0±4.4	**47.6**±6.1	**11.8**±3.8	**13.7**±4.4

Finally, a qualitative example demonstrating the functionality of the PARNorm approach is presented in Figure 6.6.

- Senior
- Male
- Upper-body clothing length short
- Upper-body clothing color gray
- Trousers
- Lower-body clothing length short
- Lower-body clothing color gray

- Adult
- Male
- Upper-body clothing length short
- Upper-body clothing color black
- Upper-body clothing color gray
- Trousers
- Lower-body clothing length short
- Lower-body clothing color black

- Adult
- Male
- Upper-body clothing length short
- Upper-body clothing color black
- Trousers
- Lower-body clothing length short
- Lower-body clothing color black

(a) Image **(b)** Baseline **(c)** Attribute-wise **(d)** PARNorm

Figure 6.6: Qualitative evaluation of the PARNorm module – An image from the test set of the Market-1501 dataset is displayed along with the predicted attributes by the baseline, with attribute-wise normalization, and finally with the PARNorm module. The utilization of only attribute-wise normalization leads to an overestimation of the recognized attributes, as two colors for the upper-body clothing are recognized. However, the application of instance-wise normalization afterward resolves that problem and yields predictions that are identical to the ground truth annotations.

The figure depicts an image from the test set of the Market-1501 dataset and the attributes recognized by the baseline approach. Additionally, the changes in recognized attributes resulting from the integration of attribute-wise and instance-wise normalization are illustrated. The predictions produced

through the employment of the PARNorm component are identical to the ground truth annotations, *i.e.*, this approach achieves perfect PAR results for this exemplary image. The baseline approach fails at recognizing the correct age and clothing colors. The age *senior* is wrongly predicted and the colors of both upper and lower-body garments, which are annotated as black, are erroneously determined as gray. Associated confidence scores are provided in Table 6.22.

Table 6.22: Comparison of normalization techniques – The table highlights the changes concerning the confidence scores achieved through the application of different types of normalization techniques. The confidence scores belong to the qualitative example in Figure 6.6.

Normalization	Age		Lower-body clothing color		Upper-body clothing color	
	Adult	Senior	Black	Gray	Black	Gray
Baseline	28.7	60.4	24.7	87.5	22.7	96.7
Attribute-wise	70.8	29.0	72.2	43.5	54.8	51.8
PARNorm	93.1	28.6	82.8	40.2	86.0	32.9

Attribute-wise normalization adjusts the output logits for each attribute and is able to correct the determination of the age and the color of the lower-body clothing. As for the color of the clothing on the torso, the use of attribute-wise normalization predicts an additional attribute, namely, predicting both the colors black and gray simultaneously. However, as mentioned in the motivation, applying instance-wise normalization afterward, fixes this overestimation issue and leads to the correct prediction of the color black instead of gray. The overestimated number of present attributes is repressed by normalizing the attribute logits for the instance. Additionally, it can be observed that instance-wise normalization enhances the classifier's certainty in its prediction when contrasted with relying solely on attribute-wise normalization. Retrieval benefits from having lower difference to binary queries obtained for these attributes. Solely using attribute-wise normalization results in confidence scores close to the decision threshold of 50% for, *e.g.*, lower-body clothing colors, which indicates limited certainty of the classifier, thus, providing less discriminating information in the attribute prediction vector used for calculating the retrieval distance.

In conclusion, the experimental validation has demonstrated the effectiveness of the PARNorm module in achieving its desired outcomes. By using this component, more balanced PAR results are obtained along with a notable improvement in retrieval performance. This applies to both cases, specializing on a single dataset, as well as generalizing to other new data sources.

6.3 Video-Based Pedestrian Attribute Recognition

In real-world scenarios, surveillance systems typically provide videos instead of single person images, providing a more extensive view of individuals over multiple time steps. As a result, entire tracks comprising multiple person crops over time become available, offering a rich source of information. One such track is shown in Figure 6.7.

Figure 6.7: Example of a person track – In contrast to single images, person tracks comprise multiple views of a person over time. The tracks are obtained by detecting and tracking individuals within videos. Tracks provide richer information about the depicted person. For instance, the handbag is only visible in certain images. Image source: [Zhe16]

The major advantage of using video-based processing is the increased amount of information available to extract the semantic attributes of a person. For instance, in the given track, the handbag is only visible in certain frames. In other frames, the handbag is occluded by the woman herself, other people, or left out of the bounding box due to inaccurate person detection. However, by aggregating information across the track, it is possible to accurately recognize the soft biometrics. In contrast, if attributes are solely recognized based on

single images, the occurrence of the person of interest might be missed during attribute-based person retrieval, as some attributes may be overlooked.

Several methods for video-based classification have emerged recently. Recurrent networks have conventionally been used for processing temporal information [Wan16]. However, incorporating specialized components such as 3D convolutions [Tra18], temporal attention [Che19], or multi-head attention [Vas17] in feed-forward networks results in even higher levels of accuracy.

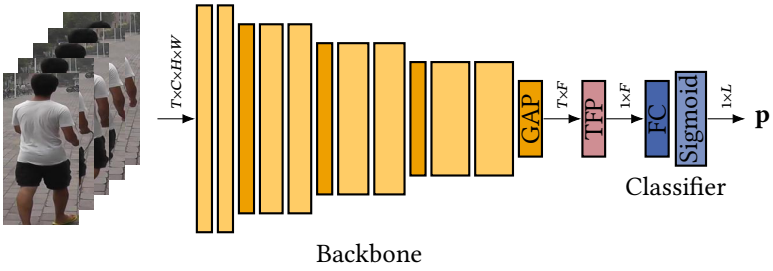
However, this thesis argues that straightforward temporal pooling through average or maximum pooling is sufficient for the PAR task. Soft biometrics, which are relevant to finding individuals matching particular personal descriptions, are independent of the movements of the person and attached objects. Despite disregarding the temporal context by ignoring the frame order, comparable accuracy is anticipated. While traits such as gait are important for identifying a specific individual in an identification or re-identification task, recognizing attributes related to clothing or an individual's age does not require such data. In addition, the temporal pooling approach offers great flexibility since the number of frames processed by the model in one forward pass can be easily adjusted. Furthermore, this approach exhibits lower computational complexity compared to the other methods considered in the evaluation. This is crucial when integrated into a complete framework with additional components such as person detection and tracking.

This section introduces temporal pooling approaches in Section 6.3.1 and thoroughly evaluates and compares them to alternative methods in Section 6.3.2. The work discussed in this section builds upon the author's study on video-based PAR [Spe20c].

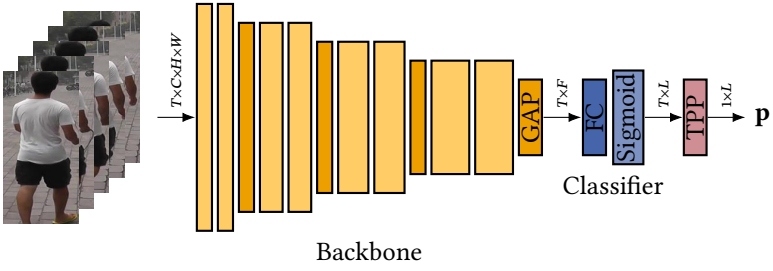
6.3.1 Temporal Pooling

In this context, temporal pooling involves combining feature representations extracted from individual person bounding boxes over a specific time period.

The pooling operation can be carried out using several methods, such as average or maximum pooling. The approaches followed in this work are visualized in Figure 6.8. Two different variants are studied which differ by the position of the temporal pooling operation. Both variants are based on a 2D CNN as the backbone model to generate global features for each bounding box of the person included in the track. With T being the number of sampled person images from the input track and C, H, and W being the image channels, height, and width, respectively, T feature vectors of size F are produced by the backbone. A short subset of length T from a track is referred to as tracklet in the following.



(a) Temporal Feature Pooling



(b) Temporal Prediction Pooling

Figure 6.8: Temporal pooling approaches –Two variants of temporal pooling, highlighted in red, are considered. Both approaches build on the feature representations generated by a 2D CNN backbone model. While the Temporal Feature Pooling (TFP) variant shown in Figure 6.8a directly pools the obtained feature vectors along the temporal dimension, Temporal Prediction Pooling (TPP) creates separate predictions for each time step and applies temporal pooling as the last step. This option is depicted in Figure 6.8b. Both approaches produce one tracklet-level attribute vector **p**.

The first variant explored involves the temporal pooling of feature representations that are produced by the backbone model. Thus, it is called Temporal Feature Pooling (TFP). As illustrated in Figure 6.8a, TFP is integrated between the spatial pooling operation of the backbone and the attribute classifier. As a consequence, the classifier generates attribute predictions based on a single track-level feature representation.

The second option creates separate attribute predictions for each of the T time steps. As a final step, Temporal Prediction Pooling (TPP) is implemented to obtain predictions for the entire track, as illustrated in Figure 6.8b.

While the second concept is the straightforward procedure, attribute predictions may be sensitive to slight disturbances in the input images and, thus, fluctuate from frame to frame. Additionally, in cases as visualized in Figure 6.7, where attribute are occluded entirely in some frames, different predictions for the same attribute are obtained. Contrary to that, global feature vectors capture the appearance of the whole body in more granularity with less semantics and may, therefore, be more robust.

Both architectures do not add any extra parameters, resulting in a lightweight model that supports fast inference. Additionally, since the models do not learn the temporal context, input tracks can be flexibly down-sampled according to the application's needs. In contrast, certain methods such as 3D CNNs and temporal attention approaches require selecting a fixed sequence length prior to training.

6.3.2 Evaluation

In this section, the temporal pooling approaches are evaluated and compared to a temporal attention approach [Che19], the 3D ResNet-50 [Har18] as a representative of a 3D convolutional model, and the VTN model [Nei21].

The evaluation is carried out on the MARS dataset [Zhe16, Che19]. For fair comparison with the 3D ResNet architecture, ResNet-50 is used as the backbone model in all experiments. During training, tracklets of length 15

are sampled from the tracks. For the temporal pooling approaches, sampling is conducted randomly to benefit from increased diversity of sampled sequences. For the other methods, consecutive frames beginning at randomly determined time steps are sampled. Furthermore, the 3D ResNet is initialized with weights pre-trained on three datasets: Kinetics-700 [Car19], MiT [Abu16], and STAIR [Yos18]. Apart from that, the training setup, schedule, and parameterization introduced for the baseline in Section 5.2.4 is utilized.

First, the impact of different temporal pooling operations is investigated. Specifically, average and maximum pooling are considered and compared in Table 6.23.

Table 6.23: Comparison of temporal pooling operations – The results achieved for average and maximum temporal pooling of backbone features and predictions are compared. The results indicate that the use of average pooling is beneficial.

Approach	Operation	MARS				
		mA	F1	mADM	mAP	R-1
TFP	Average	**77.5**	**88.2**	**70.7**	**34.8**	**35.1**
	Maximum	69.9	84.6	60.1	22.6	24.1
TPP	Average	75.7	**87.5**	**69.3**	**33.1**	**32.0**
	Maximum	**76.7**	84.2	64.2	25.8	27.3

The results demonstrate that average pooling yields superior outcomes, particularly for the TFP methodology. Maximum pooling, which concentrates on the most salient features, may be misleading when analyzing tracks. The most prominent features are often those that deviate from the rest. As a result, anomalies in individual frames of the tracklet or brief occlusions may be considered and result in inaccurate predictions. In contrast, average pooling functions as a weighted voting mechanism to ensure dominant features or predictions prevail. If occlusions or similar disturbances are temporary, average pooling mitigates the impact of those disturbances. However, for the mA used in the TPP approach, maximum pooling outperforms average pooling. The use of maximum pooling increases the positive recall of imbalanced attributes. However, it causes overconfidence in the model for these

attributes, skewing the model's output probabilities and ultimately harming attribute-based person retrieval performance. Therefore, temporal average pooling is utilized.

Next, this study compares both temporal pooling approaches to alternative video processing methods. The experimental results are presented in Table 6.24. The TFP model achieves the most favorable outcomes for both tasks and all metrics. This finding affirms the hypothesis that temporal pooling is adequate for identifying semantic person attributes, implying that explicit capture of temporal context is not necessary.

Table 6.24: Comparison of video processing methods – The results demonstrate that complex model architectures are not required for strong video-based PAR and retrieval performance, as TFP outperforms methods such as 3D CNNs [Har18], temporal attention models [Che19], and the VTN [Nei21].

Approach	MARS				
	mA	F1	mADM	mAP	R-1
3D ResNet-50 [Har18]	74.0	85.2	60.6	24.0	26.4
Temporal attention [Che19]	77.1	87.6	68.9	32.7	32.9
VTN [Nei21]	75.1	85.5	61.5	25.1	25.5
TFP	**77.5**	**88.2**	**70.7**	**34.8**	**35.1**
TPP	75.7	87.5	69.3	33.1	32.0

When comparing the two temporal pooling methods, TFP clearly outperforms TPP in both attribute-based person retrieval and PAR. In terms of PAR, especially mA differs, suggesting that pooling the probabilities for the presence of attributes often leads to predicting the absence of attributes with few training samples. Since attributes are often recognized with low certainty, averaging prediction probabilities across an entire tracklet may fail to surpass the attribute decision threshold. It is advantageous to combine global backbone features along the temporal dimension as they offer more refined information about the portrayed person. The averaged track-level feature vector appears to capture a strong representation of the person depicted in the tracklet, without omitting information that is only visible in a few frames.

In terms of the PAR results, the temporal attention model exhibits the second-best performance, with the mA showing particularly strong performance compared to the TPP approach. The attention mechanism of this model appears to concentrate effectively on the frames within the tracklets, where the attributes are clearly visible, while ignoring those frames with obstructions. Notably, the TPP approach slightly outperforms the temporal attention model in terms of retrieval mADM and mAP, which assess the entire retrieval rank lists' quality. This finding suggests that the temporal attention's enhanced focus on specific frames could negatively impact the model's calibration, leading to overconfidence, similar to the maximum pooling TPP approach.

The 3D model produces the poorest results. Such models require a significant amount of training data, especially concerning diversity for each classification category [Kat20], which is not the case for many attributes. Furthermore, the basis for fine-tuning is less adequate than for 2D CNNs, as the tasks included in pre-training datasets do not fit the PAR task seamlessly. It is assumed that the tracklets included in the MARS dataset do not provide enough data and diversity to learn robust feature representations.

Furthermore, the VTN model utilizing a transformer-based classification head also shows poor performance. This issue may stem from either insufficient training data to fine-tune the model or from the fact that the model has many design parameters that are not optimally tuned. However, experimenting with different training schedules and parameter variations did not result in significant improvement in the outcome.

The TFP approach achieves the best PAR and retrieval results across the methods. However, temporal pooling does not take into account the order of frames, unlike the other methods. As a consequence, straightforward temporal pooling is less appropriate for recognizing persons' motions or actions. This effect is demonstrated in Table 6.25, which reports the mA across the motions annotated for the MARS dataset. As anticipated, it is evident that the temporal attention approach, incorporating a designated branch for motion recognition, and the 3D model, surpass the temporal pooling approaches for

identifying individuals' movements. However, regarding attribute-based person retrieval, motion is not significant as it does not offer any substantial data for locating individuals with matching visual descriptions.

Table 6.25: Motion classification results – The mA scores for recognizing the motion in the MARS dataset are presented. Contrary to the recognition of static attributes, temporal pooling struggles to recognize people's motions and yields poor results compared to the other methods.

Approach	Motion mA
3D ResNet-50 [Har18]	68.8
Temporal attention [Che19]	**70.4**
VTN [Nei21]	67.5
TFP	68.0
TPP	64.7

Despite the fact that TFP is lightweight as it does not add any additional parameters to the baseline approach, the experimental results demonstrate the superiority of TFP for the video-based PAR and attribute-based person retrieval tasks. In order to illustrate the advantageous balance between inference time and retrieval accuracy, Figure 6.9 shows the correlation between the number of tracklets processed per second and the mADM obtained by the considered video processing methods. The average number of tracklets processed per second refers to the average over the entire test set of the MARS dataset. Processing occurs for only one tracklet at a time and frames are densely sampled. This process divides each track into sub-sequences with a length of T. Consequently, every frame within the track is processed and no sub-sampling takes place. Inference time is measured using an *NVIDIA RTX A6000* GPU with 48 GB of graphics memory. The closer the approach to the upper-right area of the plot, the better the tradeoff. The temporal pooling approaches provide the best tradeoff between computation time and retrieval performance. Especially TFP surpasses the alternative methods in both aspects. The temporal attention technique proves to be slower in the inference process than the VTN approach. However, it achieves superior outcomes in mADM. The weakest model concerning retrieval quality is the 3D ResNet model, which is the slowest as well.

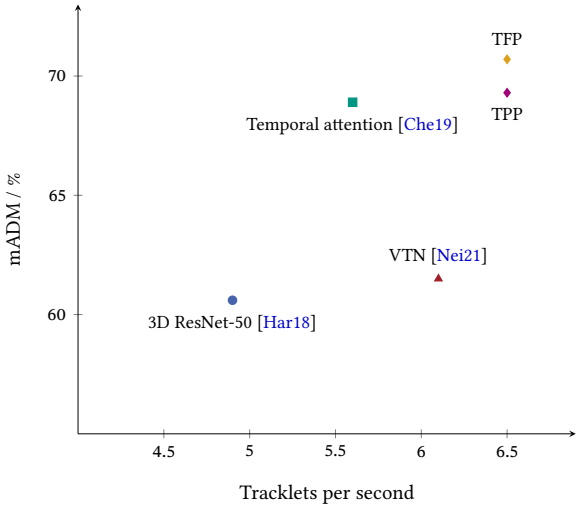


Figure 6.9: Tradeoff between inference time and retrieval accuracy – The figure illustrates the relationship between the processing speed and the mADM score achieved by several video processing methods. The closer one is to the upper-right region of the plot, the better the tradeoff. Regarding both speed and accuracy, TFP outperforms the other methods.

In conclusion, the hypothesis that simple temporal pooling is adequate for PAR with video input is valid. Moreover, pooling the feature vectors produced by the backbone model to create an overall track-level feature representation is superior to pooling the final predictions. This technique significantly outperforms all other methods in the comparison. Furthermore, temporal pooling proved to be the fastest method during inference due to its lightweight model architecture. Consequently, this approach is best suited for the use in the attribute-based person retrieval framework developed in this thesis. Nevertheless, the research highlighted that temporal pooling has weaknesses concerning the recognition of movements or actions, since the temporal context of the frames is ignored.

7 Attribute-Based Person Retrieval

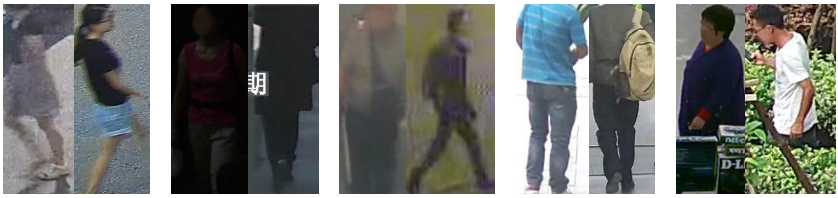
The previous chapter centered on optimizing the PAR model that functions as a feature extractor for the attribute-based person retrieval task. The soft biometric information extracted by the PAR model from person images is subsequently compared to the search query in the person retrieval stage. This chapter deals with enhancing the retrieval process itself.

For this, leveraging complementary information about the chance of failure of the PAR model is investigated in Section 7.1. Depending on the data and task-specific challenges introduced in Section 1.2, reliable recognition of all relevant attributes is not possible. For instance, some characteristics might be occluded or invisible due to poor lighting conditions. The core idea is to estimate the difficulty of attribute classification and introduce this information as a weighting mechanism during the computation of the retrieval distance.

Furthermore, three measures to ease the negative impact of skewed output distributions of the PAR model and to achieve more balanced and reliable retrieval rankings are introduced in Section 7.2. First, the application of reliability calibration is examined to align the predicted attribute probabilities with the empirical probabilities for the presence of the attributes. Differences appear due to imbalanced training distributions, overfitting, and domain gaps. Additionally, the impact of the individual attributes on the resulting ranking distance is balanced through an error-based weighting approach. Finally, an additional distance metric is proposed that takes into account the concrete output distributions of the classifier and is able to adapt to novel domains in an unsupervised manner.

7.1 Hardness Prediction

Automatic PAR encounters several challenges that are detrimental to image quality and subsequent analysis. These challenges include blurring, unfavorable lighting conditions, occlusions caused by static obstacles or other people, and potential errors in the processing pipeline, such as mislocated person detections. In the most adverse scenario, these factors make the recognition of certain attributes difficult or even impossible. Examples for this worst case are illustrated in Figure 7.1. Blurriness or poor lighting obstructs the perception of crucial details of a person's appearance resulting in uncertain or random attribute predictions generated by the PAR model. The given samples for these challenges highlight scenarios where adequate information about certain body regions cannot be gathered to reliably recognize the attributes. Similar problems arise from low-resolution images frequently encountered when processing real-world data. Additionally, images where portions of the human body are invisible prevent a PAR model from providing profound classification results for affected attributes. This problem usually stems from inaccurately aligned bounding boxes received from the person detector or occlusions. The example images depict scenarios where either the head or the lower-body regions are concealed, resulting in soft biometrics related to these areas remaining indeterminable.



Blur | Lighting | Low resolution | Misaligned bounding box | Occlusion

Figure 7.1: Example images depicting challenging factors – The example images from the RAPv2 [Li19a] dataset illustrate challenging factors that increase the likelihood of false predictions made by the PAR model.

The inability to correctly recognize attributes in such cases also causes errors in subsequent downstream tasks, such as attribute-based person retrieval. For instance, if relevant areas are obscured, the classifier may randomly select manifestations of related attributes, such as the most frequent manifestation in the training dataset, or simply assume that the attributes are absent. However, this causes serious problems in the real world. For instance, suppose the operator is searching for a person wearing gray pants. If the lower-body of the target individual is occluded, the PAR system may mistake the color of the obstructing object as the clothing color. Therefore, despite other attributes matching the query, the individual will not be considered a match and will appear low on the retrieval ranking. As a result, a relevant occurrence is missed. From the perspective of the application, it is desirable to restrict the impact of unreliable attributes to mitigate the aforementioned issue.

Existing approaches from the literature typically ignore this issue and provide classification results for all attributes regardless of the input and also consider the entirety of attributes during retrieval. Most PAR classifiers predict the presence or absence of an attribute without determining the difficulty and reliability of this prediction. The straightforward approach of assessing the difficulty of the classification task based on the classifier's confidence scores does not constitute a solid basis since very confident scores may occur even in error cases, which is shown later in this section.

Thus, an independent HP [Spe20a, Flo21] is proposed by the author of this thesis that utilizes a separate network branch to detect challenging factors and assess the difficulty of the classification task. The literature review in Section 2.3.2 demonstrates that such methods achieve promising results in a variety of applications, including object detection [Ram18], multi-class classification [Wan18a], and semantic segmentation [Rah22, Gad23]. As methods for comparison, self-referential approaches are considered. These methods estimate the certainty of attribute decisions based on the classifier's confidence scores. Sections 7.1.1 and 7.1.2 elaborate on the methodology of the independent HP and the self-referential approaches, respectively. Subsequently, gathering weights and the inclusion into the distance computation for retrieval is described in Section 7.1.3, followed by a detailed evaluation in Section 7.1.4.

7.1.1 Independent Hardness Prediction

The hardness prediction approach builds upon the foundation laid by Wang et al. [Wan18a] with their realistic predictor model for multi-class classification. Analogous to the realistic predictor, the concept of an independent HP is followed that is trained using the confidence scores of the classifier for supervision. However, to address the multi-label, *i.e.*, the multi-attribute, classification scenario and to fit the model to the specific needs arising from the application, several adaptations are employed. The most significant ones are that mid-level features are utilized for difficulty prediction to keep the model small and efficient, the abandonment of hardness feedback, and the adjusted loss function. The proposed architecture is presented in Figure 7.2, demonstrating the integration of the HP into the PAR framework.

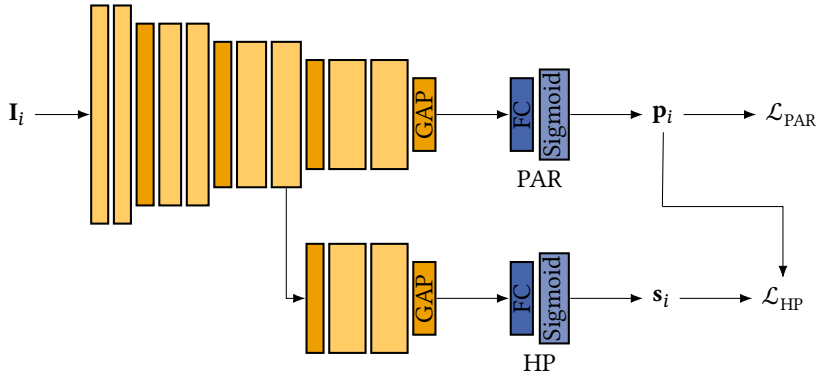


Figure 7.2: Architecture of the HP model – An additional HP branch is integrated into the PAR architecture introduced in Section 5.2. In addition to the attribute predictions $\mathbf{p}_i$, the model generates hardness scores $\mathbf{s}_i$ referring to the difficulty of classification. The HP branch is trained using a separate loss function $\mathcal{L}_{\text{HP}}$.

The upper part visualizes the attribute classifier, while the bottom branch shows the HP. The input image $\mathbf{I}_i$ is passed through the classification backbone to produce global feature representations. It is followed by the PAR head, which produces the attribute predictions $\mathbf{p}_i$ using a FC classification and sigmoid activation layer. In addition, extracted mid-level feature maps are also

processed by the hardness prediction head. To increase the number of parameters available for hardness prediction and decouple the task from PAR, the last stage of the backbone model is duplicated. The semantics of learned features increases the farther back in the CNN. Whereas PAR needs features that allow distinguishing between the absence and presence of attributes, hardness prediction aims at identifying challenging factors that are shared across the manifestations. To facilitate learning these different semantics, separating the last stage of the backbone model is beneficial. Since features like the localization of attributes is identical for both tasks, earlier splitting of the backbone is assumed to not increase performance further. After the duplicated backbone stage, a single FC layer followed by the sigmoid activation function is utilized to obtain the difficulty estimates $\mathbf{s}_i$ for the i-th image. Separate hardness scores are generated for each of the attributes, resulting in a vector $\mathbf{s}_i$ that includes one element per attribute.

The PAR branch is trained with a weighted loss function $\mathcal{L}_{\text{PAR}}$, as explained in Section 5.2.3. The outputs of the classifier $\mathbf{p}_i$ serve as the basis for calculating the hardness prediction loss $\mathcal{L}_{\text{HP}}$. The individual elements within $\mathbf{p}_i$ are interpreted as independent probabilities for the presence of the j-th attribute in the i-th image. Since the aim of the HP is to identify common obstacles that impede accurate attribute recognition, regardless of whether the attribute is present or absent in the input image, the predicted probability for the ground truth label y_{ij} is formulated as

$$p_{ij}^{y} = \begin{cases} p_{ij}, & \text{if } y_{ij} = 1 \\ 1 - p_{ij}, & \text{if } y_{ij} = 0 \end{cases}. \tag{7.1}$$

Based on this, the classification error is defined as $1 - p_{ij}^{y}$. As the HP aims at predicting the difficulty of classification, this classification error is selected as the learning target. This guarantees that high hardness scores are assigned to images for which the attribute classifier is far off the ground truth. Conversely, images that are classified correctly by the classifier with a high degree of certainty receive lower scores. Then, the loss function $\mathcal{L}_{\text{HP}}$ for training the

HP is formulated as

$$\mathcal{L}_{\text{HP}} = -\sum_{j=1}^{L}(1 - p_{ij}^{y})\log(s_{ij}) + p_{ij}^{y}\log(1 - s_{ij}), \tag{7.2}$$

which resembles the cross-entropy loss. If the classifier is certain about its prediction for the j-th attribute, meaning that ${p^{y}}_{ij} = 1$, then the corresponding loss component for this attribute is $-\log(1-s_{ij})$. This loss will be minimal if $1 - s_{ij} = 1$ applies, which requires a predicted difficulty of $s_{ij} = 0$. The higher the hardness score s_{ij} in this case, the higher the loss that is received. The loss is also minimized for incorrect predictions when the HP produces hardness scores close to 1.

Wang et al. [Wan18a] propose a hardness feedback mechanism to enhance the performance of the classifier by prioritizing difficult samples once easy samples are classified correctly. This is achieved by weighting the classification loss with the hardness scores. However, this concept carries risks, especially in the context of PAR. Certain attributes might be exceptionally challenging or even impossible to recognize, as shown in Figure 7.1. A classifier focusing excessively on such samples might achieve sub-optimal performance, as it cannot base its decision on meaningful traits. Given these concerns, the feedback from the HP to the classifier is omitted. Furthermore, no gradients from $\mathcal{L}_{\text{HP}}$ are computed for the model parameters included in the PAR branch. The reason is that the HP solves a side task and only provides additional information. This process avoids any interference with the main task and allows flexible as well as modular incorporation in any PAR model. Moreover, the entire model is trained in an end-to-end manner.

7.1.2 Self-Referential Hardness Prediction

An alternative to utilizing an independent HP is the concept of self-referential hardness prediction, where the difficulty is derived from the classifier's confidence scores. To facilitate a comparison with the suggested independent HP, three different variants of self-referential approaches are investigated.

Confidence scores: The most straightforward approach employs the predicted confidence scores directly. The classifier's certainty for its attribute predictions is determined by evaluating the proximity of the confidence scores to the decision threshold. Confidence scores close to the decision threshold indicate a high level of uncertainty about the recognition of the attributes. Assuming the default decision boundary of 0.5 for the presence of the j-th attribute, the confidence score-based failure estimate s_{ij}^{c} can be mathematically expressed as

$$s_{ij}^{\mathrm{c}} = 1 - 2 \cdot |p_{ij} - 0.5|, \tag{7.3}$$

where $|\cdot|$ indicates the absolute value. The absolute distance from the decision threshold is computed and then scaled to produce hardness scores within the interval [0,1]. Given that greater distance from the decision threshold corresponds with easier classification, the scores are inverted to measure the level of difficulty instead.

Test time dropout: Another viable approach is the injection of noise into the classifier and subsequently the quantification of the variance in attribute predictions across multiple runs of classifying the same input image. This concept enables the assessment of the robustness of classification: lower variance corresponds to less uncertainty. Dropout randomly deactivates a certain fraction r_{drop} of nodes, reducing the available information for the classification layer. It is assumed that when dropout is applied and the model is uncertain about the classification of a particular attribute, the classifier's confidence values will exhibit greater fluctuations than if the model is sure about its prediction. Given D as the number of runs and $d \in \{1, \dots, D\}$ as the dropout configuration index, the difficulty of recognition is estimated by calculating the variance of classifier outputs. Mathematically, this is represented by

$$s_{ij}^{\mathrm{d}} = \frac{1}{D} \sum_{d=1}^{D} (p_{ij}^{d} - \overline{p}_{ij})^2. \tag{7.4}$$

Here, s_{ij}^{d} represents the dropout-based hardness score for the j-th attribute in the i-th image, p_{ij}^{d} is the classifier's output for the dropout configuration

d, and $\overline{p}_{ij}$ stands for the average output of the classifier across all dropout configurations for the given sample i.

Image preprocessing: Similar to test time dropout, multiple outputs are produced for a single image to measure the robustness of attribute predictions. In this case, the ten-crop technique is employed for image preprocessing. First, images are padded with zero pixels on each side. Then, the middle as well as the four corner crops are extracted for the original and horizontally flipped image. Analogous to test time dropout, the variance across the outputs of the classifier for the ten crops are utilized as hardness estimates.

7.1.3 Weight Computation

Commonly, in attribute-based person retrieval, the similarity between a binary query attribute vector $\mathbf{q}$ and the attribute probabilities $\mathbf{p}_i$ is determined by calculating the Euclidean distance. However, this approach overlooks the reliability of classification. To address this limitation, the weighted Euclidean distance with weights obtained from hardness scores $\mathbf{s}_i$ is employed. The weighting scheme is applicable regardless of the chosen hardness estimation method, either independent or self-referential.

The weights w_{ij}^{HP} are obtained through the application of the softmax function to the inverted hardness scores as follows:

$$w_{ij}^{\mathrm{HP}} = \frac{\exp((1 - s_{ij})^{\delta_{\mathrm{HP}}})}{\sum_{l=1}^{L} \exp((1 - s_{il})^{\delta_{\mathrm{HP}}})}. \tag{7.5}$$

The hardness scores are inverted, as attributes that are simple to recognize should have a greater impact on retrieval and, thus, receive a greater weight. Furthermore, the softmax functions ensures that the resulting weights sum up to 1 for each image, which is crucial to retain comparability of distances across the gallery samples. δ_{HP} represents a temperature hyperparameter that allows the control of weight differentiation. Higher values amplify weight discrepancies among the attributes, heightening the influence of the weighting

on distance computation and, hence, the resulting retrieval ranking. On the other hand, lower values for δ_{HP} flatten the weight distribution, resulting in reduced impact of the weighting.

Finally, the weighted Euclidean distance is computed using the attribute- and sample-specific weights w_{ij}^{HP}.

7.1.4 Evaluation

In the following, the proposed HP model is evaluated. The model's training setup mirrors the baseline, using equivalent parameters, training schemes, and optimizers.

First, the proper functioning of the independent HP is validated by evaluating the hardness scores that are generated. This is achieved through analyzing images that are determined to be either particularly easy or difficult to classify. For instance, Figure 7.3 illustrates images that are considered either easy or difficult regarding the recognition of the gender.

(a) Easy **(b)** Difficult

Figure 7.3: Easy and difficult samples for gender – Images from the RAPv2 [Li19a] dataset are displayed, categorized by the independent HP as being either simple or challenging with regards to the recognition of the gender.

Easy images in this context are those where men and women are clearly identifiable due to characteristic hair length or clothing. These images have high resolution, decent lighting conditions, and well-aligned cutouts representing the whole person, which support recognition with great certainty. On the other hand, difficult samples show no clear indications for distinguishing between male and female appearance. The images exhibit low resolution, poor

contrast, multiple persons within the cutout, or no obvious gender-specific clues. Further examples for the attribute *jacket* are given in Figure 7.4. In this case, images of persons wearing clothes with short sleeves, where the skin color of the arms is clearly visible, are identified as particularly easy regarding the classification of the attribute *jacket*. Difficult images include partially visible subjects, and garments that are on the edge of being classified as other types of clothing, such as hoodies with a zipper.

(a) Easy **(b)** Difficult

Figure 7.4: Easy and difficult samples for *jacket* – Images from the RAPv2 [Li19a] dataset are displayed, categorized by the independent HP as being either simple or challenging with regards to the recognition of the attribute *jacket*.

In conclusion, the qualitative analysis suggests that the independent HP is capable of detecting challenging factors in images that increase the chance of faulty or uninformed classification decisions. These challenging factors include issues regarding the image quality as well as attributes with subjective interpretations.

However, a classifier's confidence scores can serve as a direct estimation of its uncertainty. Thus, it is crucial to examine potential differences to determine whether the independent HP is able to provide complementary information. For the analysis, confidence scores are transformed to hardness measurements, as detailed in Section 7.1.2. Afterward, instances are selected for various attributes, where the hardness scores produced by the two approaches exhibit the most substantial discrepancies. More specifically, the classifier is highly confident about its classification for the provided sample images, while the independent HP anticipates potential failure of the classifier. In essence,

the investigation explores whether the independent solution can identify challenging factors that do not result in uncertain confidence scores. The images are provided in Figure 7.5.

(a) Gender **(b)** Shoes **(c)** Hat

Figure 7.5: Misclassified samples with contradictory statements – Samples from the RAPv2 [Li19a] dataset are illustrated that are misclassified concerning the stated attribute with high confidence. However, contradictory statements about the difficulty of classification are made by the HP.

The images of individuals displayed for the *gender* attribute are erroneously classified as male. This misclassification is attributed to low resolution and the absence of crucial features such as hair and facial characteristics. Nevertheless, the attribute classifier exhibits a high level of confidence in its false predictions. In contrast, the independent HP predicts a high difficulty. It considers challenging aspects such as low image resolution and invisible body parts that aggravate the classification.

Regarding the recognition of the shoe type, the classifier predicts the absence of each possible type or the most frequent one in the training data. The relevant part lies completely outside the cropped images and, therefore, no features indicate a specific kind of shoe. Nevertheless, the confidence scores indicate high certainty of the PAR model. However, it is incorrect to assume the absence of shoes. For instance, although shoes are not visible in the cutout, the portrayed individual may still be wearing them. In attribute-based person retrieval, such samples might not be considered relevant search results, albeit depicting the individual of interest. Contrary to the confidence-based hardness estimates, the HP considers the classification of the depicted samples difficult.

Finally, two images are presented in which the attribute classifier incorrectly identifies a hat with high confidence. In contrast, the independent HP anticipates the classifier's failure accurately. A plausible explanation is that the left image has low resolution and low contrast, causing the head and hair region to blend with the background. The HP module recognizes these challenging conditions in contrast to the classifier. In the image on the right, the bun may be mistakenly identified as a bobble hat. The independent HP module appears to be aware of that risk.

The analysis shows that attribute confidences are not always dependable estimators of the likelihood of failure. This is because the model was trained to recognize attributes in all circumstances, and even when conditions permit only random guessing or reliance on prior distributions of attributes. In contrast, the independent HP is able to capture high-level factors that make the recognition task more difficult, such as poor image quality or invisible body parts. This is due to the fact that the HP is trained to recognize challenging factors through all the presence and absence of an attribute. Therefore, it can be concluded that an independent HP actually provides complementary information compared to calculating hardness scores based on the classifier's confidences.

After demonstrating the advantages of an independent HP, an ablation study is conducted to validate the design choices and proposed architecture. The findings are presented in Table 7.1. Two datasets with differing characteristics were chosen. The RAPv2 dataset includes image-level annotations and is much larger than the Market-1501 dataset, which contains attribute labels on the instance level. This is to guarantee that the design decisions are appropriate for both cases. ConvNeXt-Base served as the backbone to produce the results. First, the size of the HP branch is modified to investigate whether duplicating more or less than the last backbone stage is advantageous. The descriptions in the table denote the position of the split, *i.e.*, stage 4 indicates that the model is divided after the fourth stage of the backbone. Thus, no backbone stage is duplicated in this case. Similarly, stage 2 means that the third and fourth block of the backbone are duplicated to form the HP branch. The results provide proof that duplicating the last stage of the backbone leads to

the best results. This is especially evident for the smaller Market-1501 dataset. The difference in performance between splitting the model into two branches after stage 3 and only adding a FC layer as HP head clearly highlights the necessity of additional network capacity for the HP to produce meaningful hardness scores. For instance, the duplication of the last backbone stage increases the mADM by 1.3 points and the mAP by 1.8 points, respectively. The same tendency is observed for the RAPv2 dataset but to a much smaller extent. Enlarging the HP branch further, does not result in enhanced performance.

Table 7.1: HP ablation study – The first group evaluates varying stages for the splitting of the backbone. Splitting after stage 3 is not reported since it corresponds to the proposed approach. Furthermore, applying a positive ratio-based loss weighting function to the HP loss is examined, but results are not improved. Similarly, replacing the cross-entropy loss function with the focal or MSE loss function to train the HP does not provide any benefits. Last, providing hardness feedback to the classifier, as proposed by Wang et al. [Wan18a], is investigated but does not yield improvements.

Method	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Proposed approach	60.9	**25.7**	**18.0**	**69.1**	**35.8**	49.6
Stage 4	60.9	25.3	17.9	67.8	34.0	48.8
Stage 2	**61.1**	25.5	**18.0**	68.7	35.0	49.2
Stage 1	61.0	25.4	17.9	68.7	34.9	49.4
Loss weighting [Li15]	60.8	25.4	17.9	68.8	35.2	49.2
Focal loss	60.9	25.2	17.7	68.4	34.7	48.8
MSE loss	60.8	25.3	17.7	68.5	34.8	49.0
Hardness feedback	60.8	25.0	17.1	68.7	35.4	**50.2**

Next, the implementation of a loss weighting mechanism [Li15], analogous to that used in the PAR task, is evaluated. The loss is weighted to prioritize rarely occurring attribute manifestations over more common ones. However, there is a decrease in the obtained results. One possible explanation is that increased attention to infrequently occurring attribute manifestations leads to the HP only learning that the sole occurrence of these attributes is particularly difficult instead of identifying general challenging factors for the absences as well as presence of the attributes. Due to the low frequency in the training data, the classifier is typically less certain for such attributes and, thus, the

classification error used to train the HP is already higher, which neglects the need for additionally increasing the importance of such samples. As a result, it is proposed to build on the classical cross-entropy loss function to optimize the HP part of the model.

Two alternatives for the loss function are explored as well, namely focal loss [Lin17] and MSE loss. The findings demonstrate that, in principle, all of these losses are suitable for the task. However, regardless of the dataset, using the cross-entropy loss function achieves the strongest results concerning each of the three retrieval metrics. Last, hardness feedback to the PAR loss is investigated. The idea is to enhance the focus on difficult samples according to the HP in order to improve the recognition quality of the attribute classifier. As mentioned earlier, making use of hardness feedback is not beneficial in the context of multi-label classification with potentially fine-grained classes, since relevant clues might be invisible and, hence, focusing on such images might impede performance due to overfitting on irrelevant features.

Furthermore, it is important to understand the influence of the hyperparameter δ_{HP}, which controls the influence of the hardness-based weighting by modulating the weight differences between attributes. First, it was found that the baseline results are always outperformed by the hardness-weighted retrieval, regardless of the choice of δ_{HP} within the range of 1 to 100. This finding holds for each of the datasets. In addition, the datasets can be divided into two groups regarding the optimal choice of parameters: those with image-level annotations and those with instance-level annotations. The RAPv2 and PA-100K datasets from the first group require modest values as temperature parameters, while the PETA and Market-1501 datasets benefit from larger values and, thus, larger weight differences. This finding is attributed to the fact that the problem of invisible attributes is more severe for the datasets with instance-wise annotations, since semantic attributes that are occluded or outside the cropped bounding box are still annotated. As a result, it is difficult to choose an appropriate setting of δ_{HP} that fits all cases. However, further analysis of the hardness scores revealed that choosing the temperature parameter separately for each individual image in such a way that the resulting weights have twice the variance of the hardness scores works well across all

datasets. Since this procedure is fully automatic, there is no need to manually determine and set the hyperparameter δ_{HP}.

Table 7.2 provides the specialization retrieval results on the single datasets, *i.e.*, training and test data originate from the same data source.

Table 7.2: HP specialization results – The results demonstrate remarkable improvements through the use of the proposed independent HP.

HP	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
	61.8	24.4	24.4	68.8	26.2	34.5
✓	**63.7**	**28.0**	**27.0**	**72.4**	**32.2**	**37.2**
HP	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
	58.5	20.5	14.5	65.9	31.6	47.7
✓	**60.9**	**25.7**	**18.0**	**69.1**	**35.8**	**49.6**

The results clearly express the benefit of using the proposed independent HP to improve attribute-based person retrieval. Performance is significantly improved for all metrics and datasets. For example, the mADM, mAP, and R-1 metrics increase by 2.4, 5.2, and 3.5 percentage points, respectively, for the RAPv2 dataset. In terms of mAP, this is a relative improvement of more than 25%.

As can be seen in Table 7.3, identical observations are made for the two generalization protocols on the UPAR dataset.

Table 7.3: HP generalization results – Analogous to the specialization results, the use of the hardness scores generated by the independent HP during the computation of retrieval distances leads to a significant increase in retrieval quality.

HP	UPAR LOOCV			UPAR 4FCV		
	mADM	mAP	R-1	mADM	mAP	R-1
	56.1±5.1	17.6±4.3	19.2±8.3	46.4±5.8	11.0±3.6	12.8±4.6
✓	**58.0**±5.2	**20.7**±4.7	**21.3**±8.8	**48.1**±6.1	**13.0**±4.3	**14.2**±5.2

The results show that the independent HP is able to generalize the learned challenging factors to unseen person images from different domains. The relative improvements over the baseline reach about 17% for mAP in the generalization case.

The experimental results demonstrate great improvement by the proposed independent HP concerning attribute-based person retrieval. To further prove the advantages, a comparison with the self-referential alternatives introduced in Section 7.1.2 is performed. The results are presented in Table 7.4.

Table 7.4: Comparison with self-referential approaches – Leveraging different kinds of self-referential hardness scores also leads to improvements in terms of attribute-based person retrieval. However, the introduced HP model outperforms the alternatives for each of the metrics and datasets.

Difficulty	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	61.8	24.4	24.4	68.8	26.2	34.5
Confidence	62.5	25.4	25.2	71.9	30.4	36.4
Dropout	61.7	24.2	24.0	68.8	26.2	34.4
Ten-crop	62.1	24.6	24.2	69.3	27.2	35.3
HP	**63.7**	**28.0**	**27.0**	**72.4**	**32.2**	**37.2**
Difficulty	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	58.5	20.5	14.5	65.9	31.6	47.7
Confidence	59.7	24.3	17.3	66.7	33.5	48.8
Dropout	58.3	20.5	14.8	65.8	31.5	48.1
Ten-crop	59.4	21.4	15.5	66.4	32.3	46.5
HP	**60.9**	**25.7**	**18.0**	**69.1**	**35.8**	**49.6**

In general, the results in the table indicate that the independent HP is superior to all of the comparison methods. The self-referential approach of using the confidence scores to estimate the difficulty of classification achieves the second best results and also improves the retrieval performance on each of the datasets. Although this does not introduce any new information, increasing the focus on reliably recognized attributes and reducing the impact

of attributes with predicted confidence scores close to the decision threshold on the distance computation seems to be a promising approach to improve attribute-based person retrieval. Since this information is readily available with any PAR model, it should always be utilized to enhance retrieval performance. Measuring the fluctuations of the confidence scores over different augmentations of the input image to obtain hardness estimates also improves the retrieval quality. However, similar to using the raw confidence scores, the results are below those of the independent HP. In contrast to the other approaches compared, using dropout to produce multiple outputs for the same image does not show any improvement over the baseline. The magnitude of variation between attributes is similar and, thus, no clear discrimination is obtained in the resulting weights. As a result, the evaluation scores are similar to the baseline approach without hardness-based weighting. In summary, the results indicate the superiority of an independent and learned HP approach. The weights calculated on the basis of this approach show the best suitability for weighted Euclidean distance computation to improve attribute-based person retrieval rankings by a notable margin over the self-referential approach.

Finally, Figure 7.6 provides a qualitative comparison of the baseline and hardness score-weighted results. In each example, the first row visualizes the ranking for the baseline approach and the second row shows the improved ranking by the hardness scores of the independent HP. Blue boxes indicate the attribute query. Green and red borders denote matches and images that differ from the query by at least one attribute, respectively, according to the annotation. In each case, the top five gallery images are shown.

(a)

(b)

Figure 7.6: Qualitative comparison of baseline and HP results – In both examples, the first row shows the ranking produced by the baseline, while the second row shows the improved ranking by the use of the HP. The attribute query is represented by blue boxes. Green and red borders indicate matches and images that differ from the query by at least one attribute. The top five gallery images are displayed in each case.

Both examples display no match in the top-5 images for the baseline, but the application of the proposed hardness weighting leads to a correct image at the first rank. The first example in Figure 7.6a searches for a woman wearing a short skirt and tight pants underneath. The issue with the particular match, which is only correctly retrieved through the hardness-based weighting mechanism, is that the low resolution and contrast of the lower-body regions hardens the recognition of the lower-body clothing. Thus, the attribute classifier does not recognize the skirt, which leads to a late position in the ranking. However, the HP classifies the *skirt* attribute as difficult and, therefore, assigns a low weight to it when calculating the retrieval distance. As a result, the image is correctly ranked first, since all other soft biometrics match the query. Similar observations are made for the second example in Figure 7.6b. In this case, the person of interest's head is invisible due to a misaligned bounding box. Since the head region provides crucial information about the person's age, the classifier randomly predicts the most common age category. The HP is able to anticipate the failure of the PAR model under these circumstances. Down-weighting the age attribute results in finding the matching person image in the first position.

In summary, both quantitative as well as qualitative investigations demonstrate the effectiveness of the proposed independent HP. In contrast to self-referential difficulty estimates, the HP identifies high-level challenging factors, which ultimately yields superior attribute-based person retrieval results.

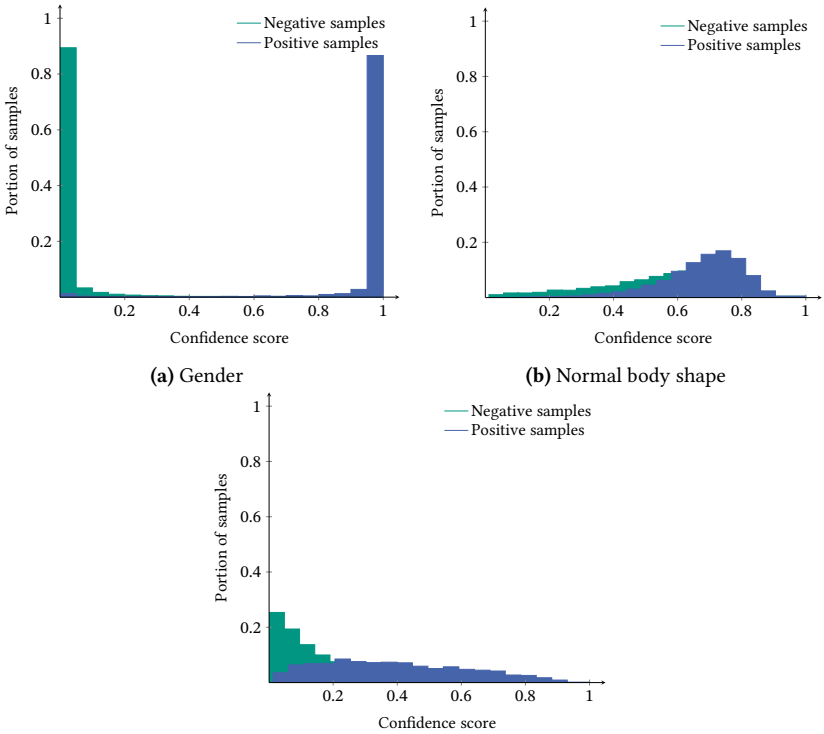
7.2 Improvement of the Retrieval Process

Achieving robust attribute-based person retrieval based on attribute predictions extracted by a PAR model requires that the confidence values closely reflect the actual probability of attribute presence based on the visual information contained in the input image. However, challenges such as suboptimal image quality, unbalanced attribute distributions, and overfitting during training often distort the outputs generated by the PAR model, causing a mismatch with the classifier's true confidence.

This discrepancy has a significant impact on the retrieval results, especially in cross-domain scenarios and real-world applications where data distributions might be significantly different. Due to a lack of appropriate training data, the classifier may struggle to confidently determine the presence of infrequent or challenging attributes on unseen data from new sources.

In Figure 7.7, output distributions of the baseline classifier trained on the RAPv2 dataset for different soft biometrics are depicted to emphasize the problem. In each plot, histograms for negative and positive samples included in the test set of the dataset are given. The histograms visualize the proportion of images for which confidence scores in a specific interval are produced.

The first example in Figure 7.7a shows the classifier's output distributions for the gender of persons. Confidence scores close to 1 indicate the presence of a woman in the input image and low output values represent males. Green and blue color stands for the distribution of confidence scores for negative and positive samples, respectively. One can observe that the model is able to reliably recognize this attribute with only few failures. Assuming the default classification threshold of 0.5, most samples are classified correctly. Additionally, the predicted scores for negative and positive samples are close to 0 and 1, which indicates great certainty by the classifier about its prediction results. The reason for the unambiguous separation between negative and positive samples is that the gender of a person can be recognized on several different visual cues. For instance, hair, clothing, or body shape provide indication about this characteristic. As a result, the classifier is able to fall back to different features, if some body parts are occluded or details are hardly visible due to poor image quality. Furthermore, both manifestations of the semantic attribute are nearly equally distributed in the training data. Thus, no bias concerning one manifestation is learned. Last, there is no significant distribution shift between training set, test set, and real-world distribution of the attribute, which also eases recognition.



(a) Gender

(b) Normal body shape

(c) Thin body shape

Figure 7.7: Distribution of attribute classifier outputs – The classifier's output distributions for three selected attributes are illustrated. Blue color visualizes the distribution for the presence of an attribute and green color denotes the absence, respectively. While the classifier reliably recognizes the gender of persons with great certainty, output distributions for normal and thin body shapes are distorted.

In contrast to this attribute, output distributions for body shapes are clearly distorted. Distinguishing body shapes based on low-resolution images is often subjective and, as a result, training data might be noisy since multiple annotators might come to different assessments for the same image. It is even possible that the labels of a single annotator might not be consistent across the entire dataset due to the tedious fine-grained annotation task and high requirements for concentration. Furthermore, the perceived body shape is

highly dependent on the clothes an individual is wearing, distance from the camera, and the body pose. Small variations may induce misleading clues and result in faulty annotations. Another important factor is an effect called regression to the mean [Spo92, Fli86], which is described in Section 2.1.2. According to it, humans tend to determine discretized attributes with fluent transitions between the manifestations as *average* or *normal*. In the concrete case, obese and thin people are disproportionately often classified to have normal body shape. Therefore, body shape attributes are imbalanced in the dataset. Almost 80% of pictures are labeled to depict a person with a normal body shape, whereas only 7% of annotations indicate thin persons. The effects on the output distributions of a classification model when trained with such imbalanced data are visualized in Figures 7.7b and 7.7c. Due to the large amount of training samples for normal body shape, the classifier is clearly superior in recognizing the presence of this attribute. Conversely, the distribution of output values for negative samples is incorrectly shifted toward predicting the presence nevertheless. The classifier is clearly biased toward predicting normal body shape for most of the images due to the dominance of such samples in the training images. Furthermore, the increased noise in the annotations additionally leads to more widespread predictions compared to the gender. Thin body shape, with few positive training samples, depicts opposite output distributions, *i.e.*, the classifier's outputs are skewed toward assuming that the person in the input image is not thin. In other words, attributes that occur seldom in training data are shifted toward lower output values and a frequently appearing one toward larger values, respectively. Discriminating between the absence and presence of the attributes is thus difficult due to overlapping distributions.

In summary, the investigation of the classifier's output distributions indicates that, although existing PAR methods [Li15, Spe23a] aim to tackle imbalanced attribute distributions during training, the challenge persists during inference and, therefore, negatively impacts attribute-based person retrieval.

In the typical attribute-based person retrieval process, the Euclidean distance is calculated between the binary query vector and the predicted probabilities for the presence of the attributes in the gallery images. The gallery images are then sorted based on these distances to form a ranked list. However, this approach overlooks the variations in the classifier's output distributions across attributes, relying solely on distances to binary values rarely produced by attribute classifiers. Consequently, some attributes exert disproportionate influence on the resulting Euclidean distances, leading to imbalanced retrieval outcomes. Moreover, variations in prediction accuracy and error rates among different attributes affect retrieval performance. Well-balanced soft biometric characteristics such as the gender are generally easier to recognize than fine-grained local attributes like glasses or imbalanced attributes, such as body shapes. Consequently, the average distance between binary query values and prediction probabilities fluctuates depending on the attribute and its manifestation.

To rectify this discrepancy and emphasize more reliably attribute-based person retrieval, three measures are proposed which address different issues, thereby promising huge potential to enhance retrieval accuracy when used in combination. Concretely, calibration, weighting, and an additional distance computation are leveraged. The methodology is based on a publication of the author of this thesis [Spe21a].

7.2.1 Reliability Calibration

Reliability calibration refers to adjusting the outputs of a model in such a way that the predicted probabilities align with the empirical probabilities observed in the data. For instance, if a well-calibrated PAR model produces prediction scores of 0.4 for a set of ten images, four images would in fact show the related attribute. The objective is to compensate for systematic biases or miscalibrations in the model's predictions. In the context of attribute-based person retrieval this is desirable, since the output scores of a calibrated model provide more reliable information for distance computation to build the retrieval rank lists.

However, due to the challenges mentioned in Section 1.2 and biased training datasets, PAR models are typically not well-calibrated, but over- or under-confident. Figure 7.8 shows the calibration curves for the baseline approach and the same attributes as given in Figure 7.7. Reliability curves are a useful visual tool for assessing the calibration of a binary classifier. These curves are created by partitioning the predicted probabilities into bins ranging from 0 to 1. Concretely, ten equally sized bins are applied in this case. On the x-axis, the average prediction score for all samples in each bin is given with the corresponding empirical probabilities on the y-axis. These probabilities are computed as the fraction of data points within the bin with a ground truth value equal to 1.

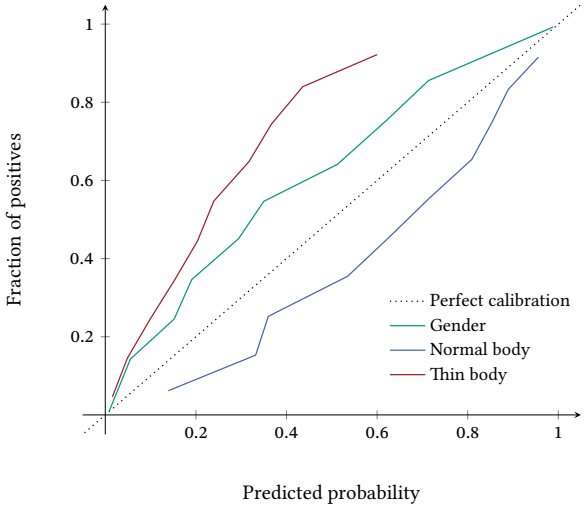


Figure 7.8: Calibration curves for selected attributes – Calibration curves for the baseline approach and three attributes from the RAPv2 dataset are shown. A well-calibrated model would produce the dotted line. If the curves are below this line, the model is overpredicting the presence of the attribute, while curves above perfect calibration indicate underconfidence.

A perfectly calibrated classifier would produce the dotted line. In this case, the predicted scores exactly match the true probabilities for an image depicting a

specific attribute. Curves above this line indicate underconfidence and curves below suggest an overconfident model. The figure shows that the baseline model is underpredicting the empirical probabilities for the gender and thin body attributes, while the model is overconfident concerning normal body shape. This finding conforms with the imbalanced distributions of the body shape attributes in the training set. Since mostly normal persons are included, this attribute is overpredicted contrary to thin people, which are rarely seen during training. Obviously, calibrating the prediction scores to align with the empirical probabilities would lead to more reliable attribute estimates and, hence, would be helpful to obtain meaningful ranking distances. If calibrated model outputs are used, the distance to the binary query in fact represents the certainty of the classifier about the presence of the attribute. As a result, the retrieved distance would be in accordance with the true probability of the image showing exactly the attributes included in the query. In the related literature, there are plenty of established approaches for reliability calibration. Thus, three different techniques from the literature are investigated in this thesis: Platt scaling [Pla99], isotonic regression [Zad01], and the spline-based approach proposed by Lucena [Luc18].

Platt scaling [Pla99] employs logistic regression. However, it assumes that the relationship between predictions and observed attribute frequencies is logistic, which is often untrue. Isotonic regression [Zad01] provides a more powerful technique that is capable of correcting monotonic distortions. The spline-based approach introduced by Lucena [Luc18] involves fitting a smooth cubic polynomial to the model's predictions and is therefore the most complex approach. The fitting of the regression models is performed on the validation splits of the datasets, which means that sufficient and diverse validation data is crucial for obtaining strong results. Without appropriate data, calibration may overfit and the calibrated model may not generalize well.

7.2.2 Error Weighting

Another issue impacting the reliability of retrieval concerns variations in average classification errors across different attributes. For instance, when

searching for a thin woman in the RAPv2 dataset, the average distances in positive matches to the binary query will be larger for the thin body attribute than for the gender, as observable in Figure 7.7. The same applies to further imbalanced or difficult-to-recognize attributes. The classifier is typically less certain in identifying these attributes, resulting in some attributes having a greater impact on the distance and, thus, the retrieval results than others.

The current retrieval approach treats each attribute equally, disregarding the differing accuracy of attribute predictions. To address this, the introduction of a weight vector $\mathbf{w}^{\text{err}}$ is proposed, which is computed as

$$\mathbf{w}^{\text{err}} = \text{softmax}(-\mathbf{m}). \tag{7.6}$$

$\mathbf{m} = (m_1, m_2, \dots, m_L)^T$ represents a vector of MSEs m_j for the attributes determined across the validation set and softmax is the softmax activation function. This allows for computing the weighted Euclidean distance and assigning weights to attributes based on their average error on the validation set. Note that identical weights are applied for each sample in the gallery.

7.2.3 Distribution-Based Distance

Even the proposed calibration and weighting mechanism are expected to not perfectly compensate for the imbalances and distorted probabilities. So, an additional and last measure is introduced, which involves fitting distribution functions to the classifier's outputs to generate additional distance values.

An ideal attribute classifier is expected to generate binary attribute predictions for person images directly. This implies that the classifier is always perfectly certain about the provided classification result. In such an ideal scenario, the Euclidean distance metric would suffice for performing robust attribute-based person retrieval. However, in practice, PAR models exhibit varying levels of confidence for images depicting a person with or without specific attributes. This variability stems from a multitude of factors, including image quality, domain gap, and the distribution and diversity of attributes within the training dataset. The influence of this factors cannot be completely

eliminated by the aforementioned mechanisms, such as calibration. Consequently, the classifier outputs a value distribution per attribute manifestation that does not necessarily peak at the extremes of 0 and 1, as was shown and discussed based on Figure 7.7.

The interplay between binary target values in the query vector $\mathbf{q}$ and the classifier's output distributions significantly affects the Euclidean distance-based rank lists. It also causes attributes to contribute to varying extents to the final distance value, depending on the attributes' manifestations. For instance, when considering Figure 7.7c, the average distance obtained for matching samples when searching for not thin people is much less than when the query includes the thin body shape attribute as a search condition. Furthermore, the distinction between matching and incorrect gallery images is hardly possible due to the overlap of distributions.

The proposed approach involves two main steps to address this challenge. First, logistic probability density functions are fitted to the classifier's output distributions for each attribute and binary value. This entails utilizing maximum likelihood estimation to obtain shifting and scale parameters for each logistic distribution. Logistic distributions are leveraged due to empirical analysis indicating their enhanced appropriateness for the task compared to other alternatives. Estimating these parameters can rely on either annotated validation data or unlabeled test data. The latter is particularly beneficial when meaningful validation data is not available or when the target domain differs greatly from the validation data in terms of attribute distributions. Since separate distributions are estimated for both the presence and absence of binary attributes, ground truth labels are required. Assuming that the attribute classifier generalizes well on the test data, the generation and use of pseudo labels, obtained by applying the default threshold of 0.5 to attribute predictions $\mathbf{p}_i$, suffice.

Based on the estimated logistic distributions, the distance component z_{ij} for the j-th attribute and the i-th image is formulated as

$$z_{ij} = \begin{cases} \mathrm{sf}_j^-(p_{ij}) + 1 - \mathrm{cdf}_j^+(p_{ij}), & \text{if } q = 1 \\ \mathrm{cdf}_j^+(p_{ij}) + 1 - \mathrm{sf}_j^-(p_{ij}), & \text{if } q = 0 \end{cases}. \tag{7.7}$$

In the equation, $\mathrm{cdf}_j^+(\cdot)$ represents the cumulative density function of the logistic distribution estimated for the presence of the j-th attribute. Similarly, the survival function of the logistic distribution determined for the absence of the j-th attribute is denoted by $\mathrm{sf}_j^-(\cdot)$. These functions' outputs represent the probabilities that the attribute is present or absent, based on the predicted confidence score p_{ij} and the actual output distributions of the PAR model. The distance z_{ij} between the query attribute value q and the i-th sample in the gallery consists of two parts. Each part evaluates the likelihood of the opposite manifestation to what the query specifies. One part is computed using the output distribution for the opposite manifestation ($\mathrm{sf}_j^-(p_{ij})$ and $\mathrm{cdf}_j^+(p_{ij})$) and, on the other hand, utilizing the respective distribution based on whether the presence or absence is specified in the query ($1 - \mathrm{cdf}_j^+(p_{ij})$ and $1 - \mathrm{sf}_j^-(p_{ij})$).

Finally, the total DBD d_i^{DBD} between the query and the i-th gallery sample is calculated according to

$$d_i^{\mathrm{DBD}} = \sqrt{\sum_{j=1}^{L} {z_{ij}}^2}. \tag{7.8}$$

This distance is always utilized in conjunction with the Euclidean distance to assure strong performance even when inadequate data for estimating the logistic distribution parameters is available. For instance, in cases where the peaks of the distributions for the absence and presence of an attribute are switched, the DBD falls back to the Euclidean distance. Note that this particular issue only applies when validation data is utilized.

7.2.4 Evaluation

This section thoroughly evaluates the retrieval optimizations detailed in the preceding sections. First, ablation studies are conducted, for instance, to identify the most suitable reliability calibration technique.

Reliability calibration: Reliability calibration methods are compared in Tables 7.5 and 7.6 for the specialization and generalization case, respectively. Regarding specialization in Table 7.5, isotonic reliability calibration of the model's outputs clearly leads to the best results. This observation is valid consistently for each of the datasets. Furthermore, improvements are also observed for the application of spline-based calibration, while Platt calibration achieves the worst performance and is even unable to improve the retrieval rankings at all in comparison with the baseline.

Table 7.5: Specialization results for reliability calibration methods – Consistently, isotonic regression performs better than the baseline and other methods. On the other hand, Platt scaling does not improve the results compared to the baseline, but the use of spline-based calibration shows an improvement.

Approach	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	61.8	24.4	24.4	68.8	26.2	**34.5**
Platt [Pla99]	61.0	23.3	23.5	66.7	24.3	33.3
Isotonic [Zad01]	**64.8**	**26.9**	**26.1**	**69.2**	**26.4**	34.0
Spline [Luc18]	64.1	26.3	26.0	69.0	26.0	33.3
Approach	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	58.5	20.5	14.5	65.9	31.6	47.7
Platt [Pla99]	56.7	19.2	13.7	65.3	30.7	45.3
Isotonic [Zad01]	**59.7**	**22.1**	**15.6**	**68.2**	**34.1**	**49.6**
Spline [Luc18]	59.0	21.3	14.9	67.8	33.7	47.1

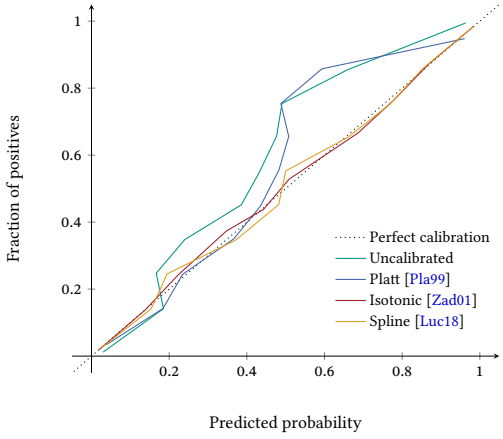
Remarkably, generalization experiments with the UPAR dataset in Table 7.6 exhibit different behavior. In this case, spline-based calibration outperforms

isotonic calibration. Furthermore, Platt's calibration technique improves the results in contrast to specialization.

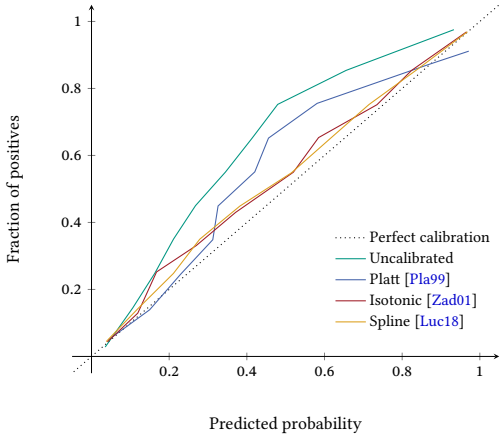
Table 7.6: Generalization results for reliability calibration methods – Contrary to the specialization results, spline-based calibration outperforms isotonic calibration in generalization. Moreover, all calibration methods lead to an increase in person retrieval performance in this scenario.

Approach	UPAR LOOCV			UPAR 4FCV		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	56.1±5.1	17.6±4.3	19.2±8.3	46.4±5.8	11.0±3.6	12.8±4.6
Platt [Pla99]	56.4±5.1	18.3±4.6	21.3±9.9	48.2±5.3	12.1±3.8	14.1±5.1
Isotonic [Zad01]	59.0±5.2	20.3±5.0	22.5±10.3	50.6±5.7	13.6±4.2	15.4±5.5
Spline [Luc18]	**59.3**±5.3	**20.6**±5.1	**23.0**±10.6	**50.8**±5.8	**13.7**±4.3	**15.7**±5.5

To delve deeper into the observation that different reliability calibration approaches are optimal for specialization and generalization scenarios, Figure 7.9 compares the calibration curves of the examined approaches for gender. The left side displays the calibration curves for the specialization scenario utilizing the Market-1501 dataset, while the right side illustrates the generalization scenario. Specifically, the fourth split of the UPAR 4FCV evaluation protocol is used for this comparison, corresponding to training the model with the RAPv2 dataset and assessing its performance with the remaining three UPAR sub-datasets. In general, the gender predictions of the uncalibrated models are underconfident in both cases. Platt scaling is able to calibrate low prediction scores but is incapable of accurately calibrating larger ones. This is particularly notable in the specialization scenario. Calibration for predictions between 0.5 and 0.7 is observed to be even worse than the uncalibrated baseline. Similar observations hold for the generalization case. However, in this case, the calibration curve is closer to the optimum, resulting in improved outcomes by using Platt's approach. A comparison of isotonic and spline-based calibrations indicates that the curves exhibit swapped progression in both instances. Specifically, isotonic calibration leads to a smoother calibration curve that is closer to the optimum in specialization.



(a) Specialization



(b) Generalization

Figure 7.9: Comparison of reliability calibration methods – Comparison of calibration curves for gender produced with various reliability calibration methods. The upper figure illustrates the specialization scenario using the Market-1501 dataset, while the lower figure compares the approaches for generalization using the UPAR dataset. One can observe that the isotonic approach achieves better calibration for specialization, whereas the use of spline-based calibration produces a smoother curve closer to perfect calibration concerning generalization.

On the other hand, the curve of this calibration method shows greater fluctuations for generalization. This discovery indicates that isotonic calibration is more susceptible to overfitting validation data compared to spline-based calibration. Although this feature has some advantages when the training, validation, and test data come from the same or similar source, it results in reduced performance in more complex generalization situations. In conclusion, isotonic calibration works well for specialization cases, while spline-based calibration is superior in generalization scenarios. Additionally, the analysis suggests that obtaining meaningful validation data is crucial to achieve a strong and robust performance of calibration methods. Without such data, calibration may become overadapted to the available data and perform poorly when applied in new domains.

Distribution-based distance: The proposed DBD is calculated by fitting logistic distributions to the classifier's outputs first. The estimated parameters may be obtained using either annotated validation data or test data from the target domain with pseudo labels. Both of these options offer distinct advantages and disadvantages. For instance, utilizing available ground truth labels in validation data may enable more precise fits of the actual distributions. Nonetheless, this procedure considers outliers and faulty labels when determining the distribution parameters. Furthermore, the collection and annotation of extensive and adequate validation data incurs significant costs. However, it remains imperative to prevent overfitting and ensure robust generalization capabilities, as demonstrated by prior investigations. Conversely, generating pseudo-labels is straightforward and cost-effective as data can be effortlessly gathered by capturing images of individuals. Additionally, the required data for the target domain can be automatically collected without any manual efforts. However, the creation of pseudo-label by simply applying the default threshold of 0.5 to attribute predictions may result in inaccurate distribution estimations that do not correlate well with the underlying distributions.

Table 7.7 presents a comparison of the supervised and unsupervised approach concerning the proposed distance measure for the single research datasets. The final results are presented, meaning calibration and error weighting

are employed as well. The only distinguishing factor is that the supervised method utilizes annotated validation data to estimate distribution parameters, while the unsupervised approach relies on pseudo-labeled test data.

Table 7.7: Supervised vs. unsupervised specialization results – The unsupervised approach that estimates the classifier's output distributions based on pseudo-labels outperforms the supervised technique that relies on validation data with ground truth labels.

Approach	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
Supervised	**65.0**	29.7	27.7	69.9	27.5	34.8
Unsupervised	64.6	**30.3**	**28.0**	**70.2**	**28.6**	**35.3**
Approach	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Supervised	60.7	25.2	18.0	69.4	36.5	51.2
Unsupervised	**61.6**	**26.7**	**19.0**	**70.8**	**37.6**	**51.9**

Unexpectedly, the unsupervised approach outperforms the supervised strategy except for the mADM metric for the PETA dataset. Using validation data with ground truth labels seems to result in overadaption. The estimated distribution parameters are too specific for the validation data and, thus, generalize poorly to the target data. Especially when only a few samples with an attribute manifestation are included in the validation set, no meaningful distribution fits are obtained. In addition, if these rare validation examples are border cases and not representative in terms of the target domain, the determined logistic distributions are less appropriate than the unsupervised ones. Unsupervised determination benefits from an increased number of samples for imbalanced attributes due to the application of a threshold of 0.5 to generate the pseudo labels. Furthermore, the samples directly originate from the target domain and, therefore, are representative. These positive effects outweigh potential disadvantages by faulty pseudo labels.

Identical observations are made for the generalization protocols on the UPAR dataset, as detailed in Table 7.8. The unsupervised approach outperforms the supervised procedure concerning both protocols and each of the evaluation metrics.

Table 7.8: Supervised vs. unsupervised generalization results – Analogous to the specialization results, the unsupervised approach proves itself superior to the use of annotated validation data.

Approach	UPAR LOOCV			UPAR 4FCV		
	mADM	mAP	R-1	mADM	mAP	R-1
Supervised	59.8±5.1	21.8±5.1	23.8±10.5	51.1±6.0	14.4±4.8	16.1±6.0
Unsupervised	**60.1**±5.5	**22.6**±5.2	**24.5**±10.5	**51.5**±5.0	**15.1**±4.8	**16.7**±6.0

Combination of approaches: Next, the impact of combining reliability calibration, error weighting, and the DBD is investigated in Table 7.9. Starting from the baseline, the approaches are added sequentially. Since weighting and computation of the DBD might benefit from calibration, calibration is applied at first, followed by attribute-wise weighting and the DBD computation.

Table 7.9: Retrieval improvements specialization results – This table evaluates the impact of the proposed enhancements concerning attribute-based person retrieval. Each of the methods demonstrates notable improvements. Combining all of them achieves the overall best results with remarkable increase in performance compared to the baseline approach.

Approach	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	61.8	24.4	24.4	68.8	26.2	34.5
+ Calibration	64.8	27.0	26.1	69.2	26.5	34.0
+ Error weighting	65.4	28.1	**27.3**	69.2	26.5	33.9
+ DBD	**66.8**	**29.1**	**27.3**	**70.2**	**28.6**	**35.3**
Approach	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	58.5	20.5	14.5	65.9	31.6	47.7
+ Calibration	59.8	22.1	15.6	68.2	34.2	49.4
+ Error weighting	60.2	23.6	16.8	69.2	35.6	49.6
+ DBD	**61.6**	**26.7**	**19.0**	**70.8**	**37.6**	**51.9**

A noteworthy finding is the dataset-dependent impact of specific enhancements on the retrieval metrics. Across datasets such as PETA, RAPv2, and Market-1501, all implemented approaches exhibit improvements when sequentially combined. However, the influence of calibration and weighting

is notably marginal on the PA-100K dataset. This is attributed to the dataset characteristics, specifically the diverse training set and the limited number of soft characteristics that are annotated, including the absence of color attributes. Such attributes are typically imbalanced and, thus, pose severe challenges in recognition. Additionally, the large and diverse training set may already lead to a well-calibrated model. The models trained on PA-100K seem to learn universal features that generalize well to unseen data. Consequently, the model demonstrates a high degree of calibration as well as similar and low attribute-specific errors, which are the basis for the proposed weighting mechanism. As a result, the computed weighting does not influence distance computation much. In contrast, employing the DBD yields clear enhancements across all metrics considered.

An overarching observation is that substantial and consistent improvements are achieved by integrating all methods for enhancement, regardless of the dataset. The greatest impact is observed in the mAP metric, indicating an optimization of the entire rank lists. By focusing on enhancing reliability of predictions and balancing distance computations among attributes, differentiation between true matches and close false positives is effectively improved. The experiments showcase improvements concerning mAP ranging from 2.4 percentage points on PA-100K to a remarkable increase of 6.2 percentage points on RAPv2. This enhancements highlight the effectiveness of the introduced optimizations in refining the accuracy and reliability of person retrieval systems across diverse datasets.

These observations transfer to the UPAR dataset for which the results are provided in Table 7.10.

Table 7.10: Retrieval improvements generalization results – The proposed measures to enhance the quality and reliability of attribute-based person retrieval exhibit great improvements regarding all metrics and both evaluation protocols.

Approach	UPAR LOOCV			UPAR 4FCV		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	56.1±5.1	17.6±4.3	19.2±8.3	46.4±5.8	11.0±3.6	12.8±4.6
+ Calibration	59.3±5.3	20.6±5.0	23.2±10.6	50.9±5.8	13.8±4.3	15.6±5.5
+ Error weighting	59.5±5.3	21.0±4.9	23.4±10.4	50.9±5.9	13.9±4.4	15.6±5.7
+ DBD	**60.1**±5.5	**22.6**±5.2	**24.5**±10.5	**51.5**±5.0	**15.1**±4.8	**16.7**±6.0

Comparing the impact of the enhancement methods, calibration and employing the DBD lead to greater improvement than the proposed weighting mechanism. Analogous to the single datasets, the combination of all methods achieves the best performance. In the generalization evaluations the mAP increases by 5 percentage points for the LOOCV and 4.1 percentage points concerning the 4FCV evaluation protocol, respectively.

In summary, the proposed techniques greatly enhance the reliability and accuracy of attribute-based person retrieval utilizing PAR models as feature extractors. The degree of improvement achieved by the single approaches thereby varies depending on the dataset-specific characteristics. Overall, combining all techniques consistently leads to superior results, on single datasets as well as in terms of the generalization experiments.

8 Evaluation

In this chapter, a comprehensive evaluation of the methods proposed within the scope of this thesis is conducted. Note that the tracking component is considered separately in Chapter 9. The chapter starts with Section 8.1, in which the integration of all methods into the unified framework, presented in Chapter 3, is examined. This evaluation encompasses the optimization of the PAR feature extractor, as introduced in Chapter 6, along with the enhancements proposed to refine the retrieval process, outlined in Chapter 7. Subsequently, Section 8.2 presents and discusses qualitative retrieval results regarding the strengths and weaknesses of the proposed framework. Following this, the combined approach is compared with current representative state-of-the-art methods from related literature in Section 8.3. The comparison includes both tasks considered in this thesis, *i.e.*, PAR and attribute-based person retrieval. Finally, Section 8.4 provides a comprehensive summary of the contents and key findings. Note that this chapter concentrates on image-based datasets due to the limited availability of representative methods for the MARS [Zhe16, Che19] dataset. The findings for MARS match those reported in this chapter.

8.1 Combination of Approaches

In the previous chapters, the proposed approaches to enhance PAR and refine attribute-based person retrieval were evaluated and discussed separately. However, in pursuing optimal accuracy, a promising strategy involves combining all the proposed methods. By doing so, one could potentially take advantage of the collective improvements brought by each individual approach, thereby further enhancing performance. The baseline approach introduced in Chapter 5 serves as basis for the investigation. Unless otherwise stated, the

hyperparameters and training scheme of the baseline approach and the individual enhancements, respectively, are leveraged. Furthermore, based on the findings in Section 6.1.2, ConvNeXt-Base is chosen as the backbone CNN for the PAR model to conduct the experiments.

8.1.1 Specialization

First, the PARNorm module is combined with the optimal design choices concerning PAR identified in Section 6.1. In the following, this best configuration of design choices is referred to as PAR optimizations. Table 8.1 reports the baseline outcomes, the performance achieved through the PAR optimizations, and the results after additionally integrating the PARNorm module.

Table 8.1: PARNorm results – The results indicate that combining the PAR optimizations with the PARNorm module improves the PAR accuracy but deteriorates attribute-based person retrieval performance. The only exception is the RAPv2 dataset.

Approach	PETA					PA-100K				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	86.1	88.1	61.8	24.4	24.4	82.2	88.5	68.8	26.2	34.5
+ PAR optim.	87.3	89.4	**68.1**	**29.4**	**27.7**	84.6	90.1	**74.0**	**31.6**	**40.6**
+ PARNorm	**88.2**	**89.5**	67.4	28.9	27.3	**84.9**	**90.2**	72.7	30.8	39.2
Approach	RAPv2					Market-1501				
	mA	F1	mADM	mAP	R-1	mA	F1	mADM	mAP	R-1
Baseline	79.3	80.0	58.5	20.5	14.5	80.7	85.7	65.9	31.6	47.7
+ PAR optim.	**80.8**	**81.1**	**61.9**	**23.6**	**17.0**	81.7	87.4	**73.2**	**40.8**	**56.0**
+ PARNorm	80.5	80.7	59.4	21.4	15.4	**84.3**	**87.5**	70.3	37.7	50.2

The findings reveal a significant enhancement in attribute recognition through the improved design characteristics of the PAR model. Combining the PARNorm component with the proposed optimizations enhances attribute recognition, except for the experiments with the RAPv2 dataset. Particularly, mA benefits while only small improvements are found for instance-based F1. These findings align with the overall objective of the PARNorm module, which aims to better balance label-based metrics, such as mA, and instance-based measures, such as F1 score. However, attribute-based person retrieval metrics deteriorate. The results suggest that the proposed

adaptations to the PAR model and training procedure interfere with learning the parameters of the PARNorm component. The primary factor is the strong regularization achieved through techniques such as SWA or dropout. Further investigations indicate that by freezing the learnable parameters in the backbone during the initial epochs, negative effects are compensated, and similar retrieval outcomes are attained when the optimizations are combined with the PARNorm module. For instance, freezing the backbone parameters for the first two epochs for the Market-1501 dataset leads to a mAP of 40.6% with nearly identical accuracy concerning PAR. Nevertheless, PARNorm is excluded from the subsequent analysis, since no further benefits in terms of attribute-based person retrieval are achieved.

Next, the impact of combining the proposed methods in terms of specialization results is evaluated. Table 8.2 provides results for sequentially adding the PAR optimizations, the independent HP, and enhancements to the retrieval process, namely calibration, error weighting, and the DBD. Since, except for the PAR optimizations, the approaches solely affect retrieval performance, only attribute-based person retrieval measures are reported in the following. Results for optimized PAR are given in Table 8.1 and in the comparison with the state-of-the-art in Section 8.3. Overall, combining the methods proposed in this thesis yields positive effects on attribute-based person retrieval for each of the datasets. Except for the mADM measure on the RAPv2 dataset and the R-1 for the Market-1501 dataset, the combination of all approaches consistently achieves the best results. The addition of each component further enhances the quality of the resulting retrieval rankings. However, the more approaches are combined, the smaller the improvements, as the proposed methods address similar goals, such as compensating for unbalanced attribute distributions. This is particularly evident in terms of mADM, which displays clear indications of saturation. This finding implies that the general appearance of the retrieval rankings to the system operator reaches the highest attainable quality through the suggested PAR method. Since mADM measures the level of agreement between annotations of retrieved images and the query, the agreement of query attributes and retrieved images at specific rank list positions remains nearly identical. The only changes that occur are images with similar agreement that switch places. This is observable from

the steadily increasing mAP and R-1 scores, which evaluate the quality of the rank lists using binary relevance labels for the gallery samples.

Table 8.2: Specialization results – The combination of the approaches proposed in this thesis results in remarkable improvements concerning attribute-based person retrieval for each of the datasets.

Approach	PETA			PA-100K		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	61.8	24.4	24.4	68.8	26.2	34.5
+ PAR optim.	68.1	29.4	27.7	74.0	31.6	40.6
+ HP	68.4	30.9	28.6	74.3	34.4	41.3
+ Calibration	69.0	32.4	29.4	74.7	35.3	43.0
+ Error weighting	**69.4**	33.3	30.5	74.7	35.3	43.2
+ DBD	**69.4**	**34.0**	**31.0**	**74.9**	**35.7**	**43.7**
Approach	RAPv2			Market-1501		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	58.5	20.5	14.5	65.9	31.6	47.7
+ PAR optim.	61.9	23.6	17.0	73.2	40.8	**56.0**
+ HP	63.7	28.0	19.8	74.5	43.2	55.4
+ Calibration	**64.4**	29.6	21.2	75.5	44.6	55.2
+ Error weighting	64.2	30.0	21.6	75.6	44.9	55.4
+ DBD	64.2	**30.4**	**21.9**	**75.7**	**45.5**	55.8

In conclusion, significant enhancements resulting from the developed approaches are evident. The absolute increases in mAP vary from 9.5 points observed for the PA-100K dataset to 13.9 points when using the Market-1501 dataset. With regard to the mADM metric, there are improvements of up to 9.8 points, observed for the Market-1501 dataset. Furthermore, the R-1 score could also be substantially boosted, albeit being subject to fluctuations due to its focus on the initial gallery sample fitting the query.

8.1.2 Generalization

Similar to the previous paragraph, the combination of the introduced approaches is examined, but in this case regarding the generalization evaluation

protocols and the UPAR dataset. Quantitative results for both protocols are reported in Table 8.3. The findings are in strong agreement with those obtained for the individual datasets. The optimum results are achieved by combining all of the suggested approaches for both evaluation protocols. Furthermore, there is also a convergence of the mADM score.

Table 8.3: Generalization results – Stacking the proposed approaches leads to notable improvement of attribute-based person retrieval. For instance, the mAP increases by 9.6 points and 8 points for the LOOCV and 4FCV evaluation protocol, respectively.

Approach	UPAR LOOCV		
	mADM	mAP	R-1
Baseline	56.1±5.1	17.6±4.3	19.2±8.3
+ PAR optim.	61.2±5.3	21.6±5.8	23.4±9.4
+ HP	62.4±5.4	24.2±6.0	24.9±9.6
+ Calibration	64.2±5.2	26.6±6.0	27.3±10.5
+ Error weighting	64.0±5.4	26.7±6.0	27.4±10.4
+ DBD	**64.3**±5.4	**27.2**±6.0	**28.0**±10.7
Approach	UPAR 4FCV		
	mADM	mAP	R-1
Baseline	46.4±5.8	11.0±3.6	12.8±4.6
+ PAR optim.	51.7±5.8	13.8±4.1	15.4±4.9
+ HP	52.8±6.2	15.4±4.9	16.4±5.6
+ Calibration	**55.6**±5.8	18.3±5.3	19.3±6.3
+ Error weighting	**55.6**±5.8	18.4±5.2	19.3±6.1
+ DBD	**55.6**±5.9	**19.0**±5.3	**19.9**±6.4

An additional noteworthy observation is that the standard deviation for mADM is almost identical for both the baseline and the optimized final result. This is especially apparent for the more demanding 4FCV protocol, which utilizes a lone data source for training in each split and the remaining UPAR subsets for evaluation. The results indicate that mADM increases uniformly across all splits, maintaining a constant performance difference between the easiest and most challenging split. In contrast, the standard deviations for mAP and R-1 accuracy significantly increase across splits as the results show

varied benefits for each split. These scores heavily rely on gallery characteristics, since these metrics are significantly affected by factors such as the number of diverse attribute sets in the gallery and the quantity of samples that agree with the queries completely. The mADM metric delivers a more comprehensive assessment of the ranking quality by accounting for partially matching gallery samples in its computation. Additionally, it incorporates normalization that accounts for biases related to the gallery and the specific query. Consequently, improvements in mADM observed while developing an attribute-based person retrieval method are expected to transfer to other domains. This is a significant advantage of the mADM in comparison to other retrieval metrics, as it simplifies the process of selecting the most appropriate model for real-world deployment.

Overall, the attribute-based person retrieval system proposed in this thesis achieves relative performance increases of 14.6% and 19.8% for the mADM using the LOOCV and 4FCV protocols, respectively. Regarding mAP, the proposed methods result in even higher relative improvement with a remarkable increase of 54.5% for the LOOCV scheme and 72.7% for the 4FCV protocol. Absolute generalization results of the LOOCV protocol which allows for more data to be used for training, reach scores comparable to the results obtained by specializing on the most challenging dataset, RAPv2. Note that this comparison is solely meaningful for the mADM metric due to increased independence from the specific queries and galleries.

To gather deeper insights regarding the influence of certain optimizations on generalization, Tables 8.4 and 8.5 provide the detailed results for each split of the LOOCV as well as the 4FCV evaluation protocol. Concerning the LOOCV scheme, the quantitative results on the four splits presented in Table 8.4 confirm the observations made using the individual datasets. Interestingly, the same decrease in mADM when adding the weighting mechanism and the DBD is observed for the split 4 as for the RAPv2 dataset. This split utilizes the RAPv2 dataset as evaluation subset, *i.e.*, this behavior is consistently observed regardless of the training data and, thus, seems to be a data-specific anomaly.

Table 8.4: UPAR LOOCV split results – The table presents the results for the individual splits of the LOOCV evaluation protocol. Information about the splits can be found in Table 4.3. Strong increase in performance is observed concerning all splits and metrics.

Approach	Split 1			Split 2		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	64.8	23.6	32.8	52.7	15.2	16.0
+ PAR optim.	69.4	28.5	38.0	58.3	18.0	18.9
+ HP	71.0	31.3	39.8	59.4	20.1	20.0
+ Calibration	72.7	34.4	44.2	60.9	22.0	21.5
+ Weighting	72.8	34.4	43.9	61.1	22.3	21.9
+ DBD	**72.9**	**35.2**	**45.3**	**61.2**	**22.6**	**22.3**
Approach	Split 3			Split 4		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	54.5	19.4	17.4	51.2	12.1	10.5
+ PAR optim.	62.0	25.6	24.1	55.1	14.1	12.5
+ HP	62.6	28.5	25.8	56.7	16.7	13.8
+ Calibration	63.8	30.3	27.4	**59.3**	19.5	16.2
+ Weighting	63.7	30.6	27.8	58.5	19.6	15.9
+ DBD	**64.3**	**30.7**	**27.9**	58.7	**20.2**	**16.6**

Moreover, the results presented in this table allow investigating and quantifying the domain gap between training using data from the same source with similar characteristics and training with diverse data from multiple sources but with divergent characteristics and underlying distributions. Note that meaningful comparison is again only supported by the mADM metric. The domain gap, measured as the relative deviation between the specialization and generalization results, narrows down when comparing the baseline results with those achieved by the framework developed in this thesis. Specifically, while the baseline approach trained and evaluated on the PETA dataset outperforms the results achieved by training the same model using the other three datasets (split 3) by 13.4% in mADM, the superiority is only 7.9% when comparing the optimized approaches. This refers to a reduction of the absolute difference in mADM from 7.3 to 5.4 points. Analogous, the relative deviation decreases from 30.6% to 22.4% when evaluating on the PA-100K

dataset (split 2) and from 14.3% to 9.4% for the RAPv2 dataset (split 4), respectively. The only exception is the Market-1501 dataset (split 1). In this case, the superiority increases when the optimized approaches are compared. The reason is that this dataset includes limited diversity and severe biases. Therefore, overfitting to the training data with similar characteristics is rewarded. In conclusion, the proposed framework for attribute-based person retrieval not only significantly improves the retrieval results, but also mitigates the impact of the domain gap between the training and test data on the retrieval performance.

Contrary to the LOOCV evaluation protocol, the 4FCV protocol is more difficult as it uses only one dataset for training in each split and then assesses the performance with the remaining three UPAR sub-datasets. The results for the individual splits are outlined in Table 8.5. Note that in this case, the standard deviation reflects the deviation of metrics across the three evaluation datasets. In general, the observations are consistent with the specialization case and those observed for the LOOCV protocol.

Table 8.5: UPAR 4FCV split results – The table presents the results for the individual splits of the 4FCV evaluation protocol. Information about the splits can be found in Table 4.3. Strong increase in performance is observed concerning all datasets and metrics.

Approach	Split 1			Split 2		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	38.3±2.2	6.4±1.9	5.9±1.9	54.6±5.7	16.4±4.8	18.5±9.3
+ PAR optim.	43.8±2.4	8.7±2.2	8.3±2.7	60.1±5.8	20.2±5.8	22.0±9.9
+ HP	44.0±2.5	9.3±2.2	8.4±2.3	61.4±5.7	22.8±5.8	23.8±10.1
+ Calibration	**47.3**±2.2	11.5±2.3	10.1±2.2	63.6±5.6	25.8±5.8	26.9±10.8
+ Weighting	47.2±2.3	11.6±2.4	10.3±2.3	63.5±5.6	25.8±5.6	26.6±10.3
+ DBD	47.2±2.2	**12.0**±2.5	**10.5**±2.3	**63.8**±5.5	**26.4**±5.5	**27.6**±10.9
Approach	Split 3			Split 4		
	mADM	mAP	R-1	mADM	mAP	R-1
Baseline	46.8±3.8	9.7±2.2	12.1±5.9	45.7±2.9	11.3±0.6	14.7±4.8
+ PAR optim.	52.4±3.9	12.7±2.5	14.8±6.4	50.4±2.2	13.5±1.5	16.6±5.4
+ HP	52.9±4.1	13.6±2.6	15.0±6.7	53.0±2.4	16.0±1.7	18.5±5.9
+ Calibration	**55.0**±4.5	15.9±3.3	17.4±7.5	**56.6**±2.7	19.9±2.7	22.7±7.2
+ Weighting	**55.0**±4.5	16.0±3.2	17.4±7.2	**56.6**±2.7	20.2±2.6	22.7±7.0
+ DBD	54.8±4.5	**16.8**±3.3	**18.2**±7.8	56.5±2.7	**20.9**±2.1	**23.3**±6.9

Using the Market-1501 dataset for training (split 1) leads to the worst results. This strengthens the finding that this dataset is notably biased and lacks diversity, resulting in inadequate suitability for training a generalizable model. Additionally, the results indicate that increased training data is beneficial for generalization, since this typically correlates with heightened diversity. The most optimal generalization results are achieved through the utilization of the PA-100K dataset for training, which is the largest of the sub-datasets included in the UPAR dataset, followed by the second largest dataset RAPv2 (split 4). However, it is striking that the difference in mADM between training with the small PETA dataset (split 3) and the much larger RAPv2 dataset (split 4) is small for the optimized approach, or even swapped when the baseline results are compared. This is due to the increased diversity of scenes and individuals within the PETA dataset, which compensates for its significantly lower number of training images when compared to the RAPv2 dataset. The PETA dataset comprises ten sub-datasets with diverse characteristics, whereas the RAPv2 dataset only contains indoor scenes, tightly aligned bounding boxes, and mainly good illumination from artificial lighting. This highlights that having diverse training data is equally crucial in achieving strong generalization outcomes for attribute-based person retrieval as having a large amount of data available for training.

8.1.3 Inference Time

Figure 8.1 illustrates the time taken per person image by the proposed approach for extracting the semantic attributes for the Market-1501 test set. The inference time measurement includes data loading, model forwarding, and storing of the respective outputs to obtain meaningful insights for real-world applications. The processing framework utilized in this experiment is *PyTorch*[1]. As the GPU, an *NVIDIA GeForce RTX 3090* is employed. It is installed in a server with 256GB of RAM and an *Intel Xeon Silver 4210R CPU @ 2.40GHz*. The achieved inference times and corresponding mADM scores are provided for multiple batch sizes B.

[1] https://pytorch.org/

The results indicate no increase in inference time by the optimization of the PAR model, as no additional parameters are added. In contrast, the inference time rises by about 2.5 ms when integrating the HP and considering batches of size 1. Notably, the observed increase in processing duration vanishes for larger batch sizes of 16 or 64. Additionally, the figure shows that the differences in inference time between larger batch sizes are minimal. Thus, the results of further experiments, which are not reported, do not prove benefits from continuing to increase the batch size.

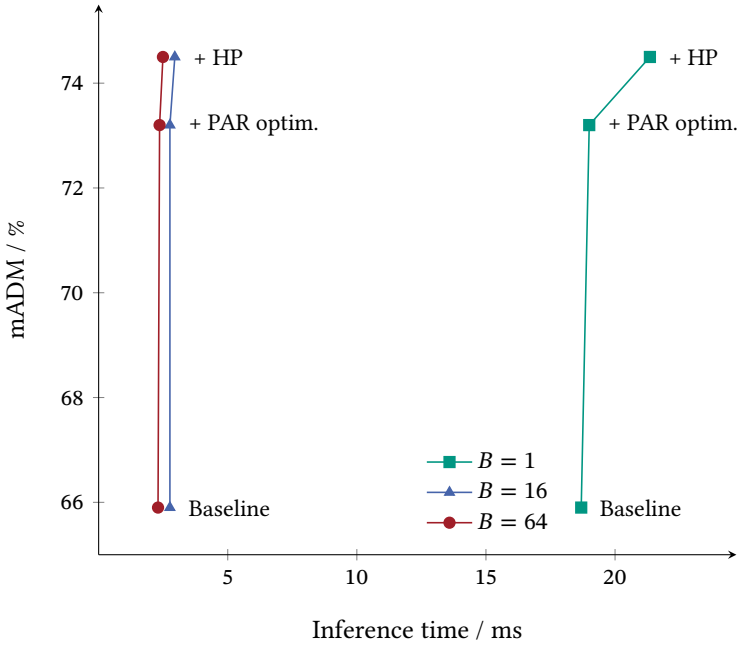


Figure 8.1: Inference times for the proposed model – The inference times per person image for the combination of the proposed model adaptations and respective mADM scores for the Market-1501 dataset are visualized. One can observe that the PAR optimizations do not lengthen the inference times. The increased inference time caused by the incorporation of the HP branch vanishes for larger batch sizes B. The model is based on ConvNeXt-Base as the backbone and measurements are conducted with a *NVIDIA GeForce RTX 3090* GPU.

In conclusion, the experiments demonstrate that the proposed feature extractor for attribute-based person retrieval is able to efficiently process a large amount of person images in a reasonable time frame. For batch sizes of 16 or larger, only a negligible rise of inference time compared to the baseline is found. Specifically, the processing duration per image is 2.5 ms for batches of size 64. This corresponds to extracting soft biometric characteristics for more than 400 individuals per second, which is assumed to be adequate for real-world applications.

8.2 Qualitative Evaluation

After presenting the quantitative improvements accomplished by the proposed framework in the previous section, this section conducts qualitative evaluations and comparisons of attribute-based person retrieval results. Initially, the investigation focuses on the specialization scenario before extending to the generalization case. In each example, the blue box denotes the query attributes, *i.e.*, the description of the visual appearance of the individual that is searched. Right of the query, the top-5 retrieved gallery images are shown. Thereby, a green border denotes a perfect match with the query description based on the ground truth annotations, whereas a red border indicates a gallery sample showing a person with a deviating set of attributes.

8.2.1 Specialization

The following figures present examples from the Market-1501 [Zhe15, Lin19] dataset, as it offer a wide range of challenges. In the initial example, depicted in Figure 8.2, the query describes an adult woman wearing green pants, a short and black upper-body garment, and carrying a handbag. Upon examining the baseline result in Figure 8.2a, it is observable that the approach encounters difficulties regarding several attributes. Although the first-ranked image appears to be a match, there are discrepancies in lower-body clothing color, gender, and upper-body clothing for the other samples.

(a) Baseline

(b) Proposed framework

Figure 8.2: First qualitative specialization result – In contrast to the baseline approach, the proposed framework retrieves five matching images. The figure shows an example for the Market-1501 [Zhe15, Lin19] dataset. The blue boxes include the query attributes and green and red borders stand for correctly retrieved images and images deviating from the query, respectively.

In contrast, the proposed framework retrieves matching samples for all of the top-5 ranks. Multiple findings are of interest. First, the baseline approach is not able to retrieve very similar input images at adjacent positions. Specifically, under the proposed framework, images ranked three to five exhibit similar appearances with only minor differences in lighting and background. Despite this, the baseline method only ranks one of these images in first position and disregards the rest. This finding reveals that the attribute predictions generated by the baseline model are highly sensitive to even slight variations in input images. The proposed framework fixes this issue through the PAR optimizations and robustly assigns comparable predictions to similar images. As a result, these samples are closely ranked, greatly enhancing the reliability

of the rankings. Furthermore, differences in color perception due to different cameras and lighting conditions are apparent in Figure 8.2b. Whereas the trousers of the first two ranked images appear to be green as specified in the query, the color seems to be more of a blue tone in the further samples, albeit all samples showing the identical person and trousers. This poses a significant challenge to the approaches but also to manual investigations. The proposed approach can deal with this challenge and retrieves the correct person regardless of the camera and corresponding appearance of colors. The reasons are two-fold. First, contrary to baseline, the measures to compensate for imbalanced attributes such as colors increases the positive recall and, thus, the optimized model is more certain about its predictions of green lower-body clothing color. Second, the introduced enhancements in terms of attribute-based person retrieval, particularly calibration and the DBD, further reduce the impact of the underconfidence regarding such attributes. In addition, the proposed framework ranks the images in which the pants appear more green in earlier positions than those with bluer perception. It is able to capture those fine deviations and reflect them in the predicted confidence scores. So, the resulting rankings are reliable and the sorting of the gallery samples aligns with human expectations and perception.

Another example is provided in Figure 8.3. In this case, the baseline approach retrieves images in early ranks that agree with the query except for the hat. This is due to the few samples of persons wearing a hat in the training data. The positive ratio of this attributes is only 2.9%. Consequently, the baseline model only recognizes 8.2% of the hats in the test set. This results in attribute predictions that do not differentiate between gallery samples, as the confidence scores of the images are similar whether or not a hat is present. Considering the methodology introduced in this work, it is able to correctly identify 73.1% of the hats, therefore allowing to distinguish persons with and without hats during retrieval. As Figure 8.3b demonstrates, each of the images ranked at the top-5 positions depict an individual matching the query description.

(a) Baseline

(b) Proposed framework

Figure 8.3: Second qualitative specialization result – The proposed framework is able to retrieve five matches for the query, whereas the baseline approach provides images of persons without hats. The figure shows an example for the Market-1501 [Zhe15, Lin19] dataset. The blue boxes include the query attributes and green and red borders stand for correctly retrieved images and images deviating from the query, respectively.

However, there are a few queries for which the proposed framework achieves worse retrieval results in mADM than the baseline approach. All of them show similar issues. A representative example is visualized in Figure 8.4. In this example, the baseline achieves good results with four matches within the first five samples of the ranking. In contrast, the optimized framework only includes a single match. One can observe that the difference between the query and the incorrectly retrieved individuals in the early ranks is the bag. Apart from that, the semantic attributes agree with each other. Delving into the reasons for this deterioration, it is found that the issue stems from the

HP approach. The attribute classifier is uncertain about bags, especially concerning differentiation between the categories *handbag* and *bag*. As a result, the independent HP considers the classification of these attributes difficult for the given samples since only the strap is visible. This, in turn, leads to minor impact on the retrieval distance and, hence, to a low distance to the query. However, since this only affects very few cases negatively but is beneficial in much more queries, this does not represent a severe issue.

(a) Baseline

(b) Proposed framework

Figure 8.4: Third qualitative specialization result – In this example, the proposed framework erroneously retrieves images of persons with a bag in early positions, albeit a bag is not specified in the query. The figure shows an example for the Market-1501 [Zhe15, Lin19] dataset. The blue boxes include the query attributes and green and red borders stand for correctly retrieved images and images deviating from the query, respectively.

8.2.2 Generalization

Similar to the specialization scenario, the proposed framework's generalization performance is evaluated by comparing it with the baseline results. To accomplish this, split 2 of the UPAR LOOCV evaluation protocol is employed. Specifically, the model is trained using the PETA [Den14], RAPv2 [Li19a], and Market-1501 [Zhe15, Lin19] datasets, and the evaluation is conducted using the image data from the PA-100K [Liu17] dataset. The first example is illustrated in Figure 8.5. Adult males with short hair, short and red upper-body clothing, and green trousers are searched.

(a) Baseline

(b) Proposed framework

Figure 8.5: First qualitative generalization result – The figure shows a generalization example for the UPAR dataset. The blue boxes include the query attributes and green and red borders stand for correctly retrieved images and images deviating from the query, respectively. In contrast to the baseline, the proposed framework correctly recognizes the green trousers of the person of interest. Source of the images: [Liu17].

Analyzing the ranking list generated by the baseline, it is apparent that the retrieved images correspond with the person description with the exception of the color of the pants. In the first five ranks, this method does not find any matches. On the contrary, the proposed framework achieves perfect results as it provides the sole match included in the gallery for this query at the first position. This improvement is credited to improved PAR under demanding scenarios and imbalanced attribute distributions. The matching sample suffers from low spatial resolution and few samples with green trousers in the training set, resulting in uncertain predictions produced by the baseline approach. The optimized model robustly extracts semantic attributes despite these difficulties. In addition, the refined attribute-based person retrieval distance handles varying output distributions across attributes, particularly related to imbalanced attributes. As a result, the correct image showing the rare green lower-body clothing is ranked earlier than those with other lower-body clothing colors.

A further example is shown in Figure 8.6. In this case, the images retrieved from the gallery by the baseline method do not match the query based on different attributes, including the color of the torso, gender, or backpack. Consequently, the ranking seems to be unreliable. The baseline method struggles with transferring its learned features to new data sources. In contrast, the proposed framework retrieves images of the correct person from the gallery database. The employment of regularization techniques, such as SMA, proves to be effective in preventing overfitting. Hence, the proposed method is able to learn universal features that generalize to data originating from various scenarios. Moreover, it is remarkable that the person of interest is retrieved even when the backpack is not or scarcely visible (ranks 3 and 4). This highlights the power of the proposed HP and underscores its capability to generalize. Besides, it is worth noting that the fifth retrieved image displays the individual of interest but does not qualify as a match due to the invisible head. The image-wise annotations for this image lack information about the hair, *i.e.*, no hair length is specified. Consequently, the annotations differ from the query concerning the short hair attribute. Nonetheless, the HP detects this and reduces the impact of related attributes during the retrieval process, resulting in the image being retrieved at an early position. However, this

samples is still ranked lower than the others. The ranking generated is plausible and reliable as images clearly showing all query attributes appear first in the ranking, followed by those where some attributes are barely visible or completely invisible. It can be concluded that the distances produced by the proposed framework measure the similarity to the query in accordance with human expectations.

(a) Baseline

(b) Proposed framework

Figure 8.6: Second qualitative generalization result – The figure shows a generalization example from the UPAR dataset. The blue boxes include the query attributes and green and red borders stand for correctly retrieved images and images deviating from the query, respectively. The proposed framework is able to retrieve gallery samples at early ranking positions even when relevant attributes, such as the backpack, are invisible. Source of the images: [Liu17].

Last, exemplary cases for poor retrieval returns by the proposed framework are presented in Figure 8.7. The example in the uppermost row implies difficulties concerning openly defined semantic attributes. The *other* category of clothing colors includes various clothing patterns lacking a distinct primary color and colors that are non-listed as separate category. Consequently, the model experiences difficulties in learning resilient features necessary for recognizing these attributes. Additionally, the intra-class variation of these attributes found across different datasets is immense and impairs the model's generalization capability. Such issues lead to the ranking without a match among the top-5 positions. Patterns on clothing are only recognized for either the upper- or lower-body, but not both as specified in the query.

The example in the second row illustrates a scenario where most of the semantic attributes of the gallery samples retrieved match those of the query. However, minimal variations concerning single attributes are exhibited. For instance, the images at ranks 1 and 5 do not display a bag. Similarly, in the second and fourth image, the person wears wide trousers instead of a skirt which creates confusion. These findings suggest potential for improvement in further optimizing the accuracy of PAR, specifically mA. The retrieved images are comparable to the query in terms of overall appearance (reflected by the instance-wise F1 score). However, issues concerning details are noticeable in this example (related to mA). This finding hints that it may be beneficial to conduct further research to improve mA without adversely affecting retrieval performance or F1. This corresponds to the aim of the PARNorm module proposed in Section 6.2.

The last illustration in the third row indicates that external factors such as lighting or noise still impact the generalization outcomes achieved by the proposed framework. The images retrieved at positions two to four exhibit a gray haze, which leads to clothing colors being classified as gray even if they are not the true colors according to the ground truth annotations. This finding indicates that to effectively address the problem of strong generalization under all conditions, it is necessary to utilize large and diverse datasets, such as the UPAR dataset. However, gathering ground truth annotations in this

case is challenging, as it is also difficult for human annotators to accurately determine the true clothing color in such images.

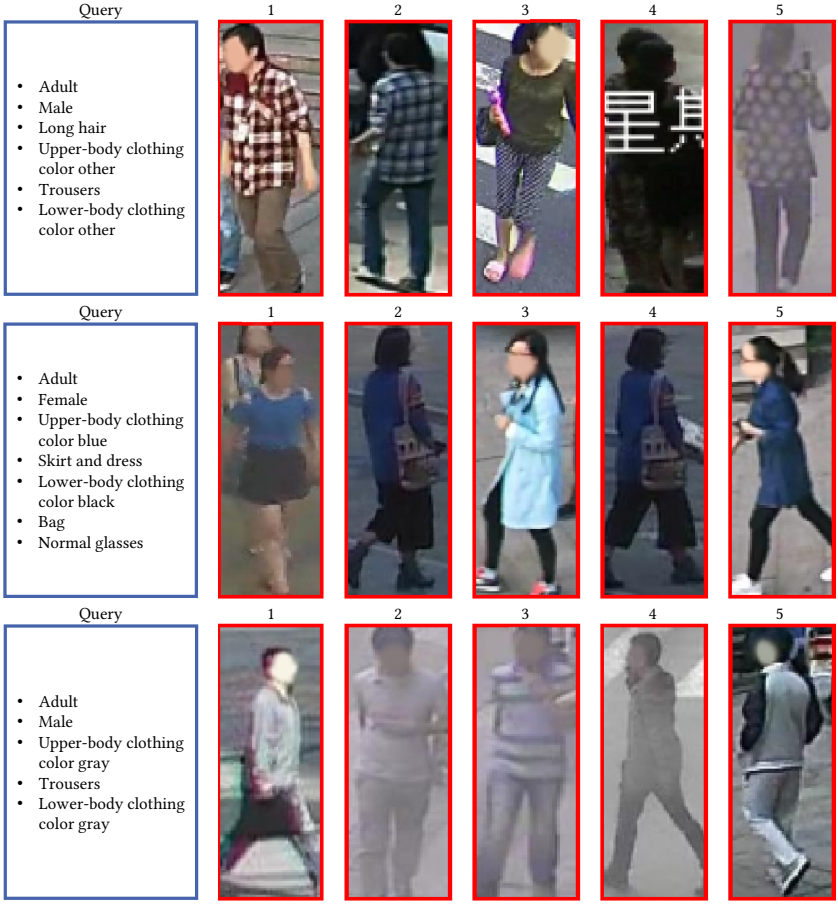


Figure 8.7: Analysis of error cases – Three representative generalization examples from the UPAR dataset are provided for which the proposed framework has difficulties in retrieving a match within the early ranks. In the top row, the problem is related to the openly defined color category *other*. The second row depicts images that are similar to the query but not match entirely. In the last row, poor image quality results in faulty color predictions. Source of the images: [Liu17].

In summary, the qualitative analysis of attribute-based person retrieval results highlights the remarkable improvements observed in the quantitative analysis. The proposed methodology is effective in generating strong and reliable retrieval outcomes. Moreover, investigations reveal that the approach generalizes well to unseen data sources. This applies to both the PAR model and the HP, as well as the retrieval optimizations proposed. The examination evidenced that widely defined attributes, imbalanced and fine attributes, and noisy images still pose challenges, ultimately impacting the quality of retrieval rankings.

8.3 Comparison with the State-of-the-Art

In this section, the results obtained by the proposed method are compared to the state-of-the-art approaches from the related literature. First, it is focused on the specialization scenario on the individual datasets. The comparison is conducted for PAR as well as the attribute-based person retrieval task. Afterward, the same analysis is performed for generalization on the UPAR dataset.

Note that there are gaps in the results as approaches from the literature often only evaluate on a limited number of datasets or apply inconsistent evaluation protocols. To ensure a fair comparison, solely approaches are considered that evaluate using identical splits and evaluation protocols.

8.3.1 Specialization

Table 8.6 reports the state-of-the-art results concerning PAR for the PETA [Den14], PA-100K [Liu17], RAPv2 [Li19a], and Market-1501 [Zhe15, Lin19] datasets.

The proposed approach yields state-of-the-art results, although its development focused on attribute-based person retrieval optimization. Specifically, the proposed framework, utilizing ConvNeXt-Base as the backbone, outperforms all other methods in terms of instance-based F1. When it comes to

mA, more complex approaches surpass the method for the PETA and PA-100K datasets. Note that the PARFormer [Fan23] applies different backbone architectures, which complicates the comparability. Similar findings are observed for the proposed framework with ResNet-50 as the backbone. Other approaches leveraging the same backbone model are unable to achieve better results in terms of F1. Overall, it is impressive that the straightforward and lightweight architecture utilized in this thesis achieves state-of-the-art performance and surpasses more complex methods from the literature. This demonstrates that concentrating on the thorough analysis and mitigation of the most severe challenges in PAR, as was done in Section 6.1, can be superior to developing novel modules that add numerous parameters to the models. In particular, for applications where inference time requirements are a crucial issue, this approach proves advantageous.

Table 8.6: State-of-the-art PAR – The proposed framework achieves state-of-the-art results in PAR even though it is designed and optimized for attribute-based person retrieval. **Red** and **blue** colors denote the best and second-best results, respectively. † Results were produced using the official implementation.

Method	Backbone	PETA		PA-100K		RAPv2		Market-1501	
		mA	F1	mA	F1	mA	F1	mA	F1
MsVAA[Sar18b]	ResNet101	84.6	86.5	–	–	–	–	–	–
VAC [Guo19]	ResNet-50	83.6	86.2	79.0	86.8	–	–	–	–
ALM [Tan19c]	BN-Inception	86.3	86.9	80.7	86.5	–	–	–	–
JLPLS-PAA [Tan19b]	SE-BN-Inception	84.9	86.9	81.6	87.3	–	–	–	–
JLAC [Tan20]	ResNet-50	87.0	87.5	82.3	87.6	–	–	–	–
MSCC [Zho21]	ResNet-50	–	–	82.1	86.8	**80.2**	79.1	–	–
SB [Jia21b]	ResNet-50	84.0	86.4	80.2	87.4	78.5	78.7	–	–
SB [Jia21b]†	ResNet-50	84.0	86.3	81.6	88.1	77.4	78.5	76.5	83.6
SSC_{soft} [Jia21a]	ResNet-50	86.5	87.0	81.9	86.9	–	–	–	–
MCFL [Zhe21]	ResNet-50	86.8	86.7	81.1	87.4	–	–	–	–
Rein-PAR [Ji22]	ResNet-50	85.5	85.9	80.6	85.7	–	–	–	–
DRFormer [Tan22]	ViT-Base	90.0	88.3	82.5	88.0	–	–	–	–
VTB [Che22b]†	ViT-Base	86.8	86.9	83.6	88.0	78.5	79.8	**80.8**	**85.2**
COB [Zho23]	ResNet-50	86.4	86.8	84.5	87.0	–	–	–	–
	ConvNeXt-Base	88.1	88.5	**88.1**	**89.1**	–	–	–	–
PARFormer [Fan23]	Swin-Base	**88.7**	88.7	84.0	87.7	–	–	–	–
	Swin-Large	**89.3**	**89.1**	84.5	88.5	–	–	–	–
Proposed framework	ResNet-50	87.1	87.7	82.2	88.5	79.4	**80.1**	79.4	85.1
	ConvNeXt-Base	88.2	**89.5**	**84.9**	**90.2**	**80.5**	**80.7**	**84.3**	**87.5**

Next, the state-of-the-art results regarding the primary task addressed in this thesis, namely attribute-based person retrieval, are presented in Table 8.7. Scores for mADM are not reported since there are no values for comparison yet.

Table 8.7: State-of-the-art attribute-based person retrieval – The proposed framework achieves state-of-the-art results for each of the datasets and regardless of the backbone model. The only exception is the R-1 metric on the Market-1501 dataset. When two backbone models are stated, the first one represents the image, while the second one represents the text or attribute encoder. Multi-layer Perceptron (MLP) refers to a small network of FC layers. Typically, three to four layers are used. **Red** and **blue** colors denote the best and second-best results, respectively. †Results were produced using the official implementation.

Method	Backbone	PETA		PA100K		RAPv2		Market-1501	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
AAIPR [Yin18]	ResNet-50 + MLP	–	–	–	–	–	–	20.7	40.3
AIHM [Don19]	ResNet-50 + Word2Vec [Mik13]	–	–	–	–	–	–	24.3	43.3
SAL [Cao20]†	ResNet-50 + MLP	–	–	15.0	22.7	–	–	29.4	44.4
TAVD [Iod20]	ResNet-50 + GloVe [Pen14]	–	–	18.2	33.8	–	–	34.4	46.1
SB [Jia21b]†	ResNet-50	20.5	20.7	24.3	33.5	17.5	12.0	25.5	37.8
SB [Jia21b]†	ConvNeXt-Base	24.4	24.4	26.2	34.5	20.5	14.5	31.6	47.7
ASMR [Jeo21]	ResNet-50 + MLP	–	–	20.6	31.9	–	–	31.0	49.6
ASMR [Jeo21]†	ResNet-50 + MLP	–	–	–	–	–	–	30.1	43.2
MCML [Zhu23a]	ResNet-50 + MLP	–	–	–	–	–	–	36.4	**52.5**
Proposed framework	ResNet-50	**25.6**	**24.4**	**29.1**	**34.8**	**26.4**	**18.6**	**38.0**	52.1
	ConvNeXt-Base	**34.0**	**31.0**	**35.7**	**43.7**	**30.4**	**21.9**	**45.5**	**55.8**

The proposed method with ConvNeXt-Base as the backbone achieves superior results on all datasets and metrics. The second-best performance is also achieved using the proposed framework, but with ResNet-50 as the backbone model. The only exception is the Market-1501 dataset, where the obtained R-1 accuracy with ResNet-50 slightly falls short of the score achieved by the MCML [Zhu23a] method. However, in terms of the more important mAP measure that evaluates the quality of the overall rankings, the proposed framework demonstrates clear superiority. The proposed framework surpasses the best performing method from the literature, namely MCML [Zhu23a], by impressive 1.6 and 9.1 points in mAP depending on the backbone used.

8.3.2 Generalization

The following study compares the proposed framework with existing literature approaches for both generalization evaluation protocols of the UPAR dataset. Fewer methods are compared than in the specialization scenario due to the need for re-training of approaches with the new UPAR dataset. The unavailability of official implementations and problems with reproducing the results of works from the literature through re-implementations hinder the reporting of reliable results for additional methods. For fair comparison, such approaches are excluded and only reliable results are presented.

First, the state-of-the-art results for the LOOCV evaluation protocol are investigated in Table 8.8. This protocol involves training with three sub-datasets in each split and evaluation on the fourth one.

Table 8.8: State-of-the-art UPAR LOOCV – Generalization results for the UPAR LOOCV protocol. The proposed framework outperforms the methods from related literature, particularly concerning the attribute-based person retrieval task. When two backbone models are stated, the first one represents the image, while the second one represents the text or attribute encoder. MLP refers to a small network of FC layers. Typically, three to four layers are used. **Red** and **blue** colors denote the best and second-best results, respectively. † Results were produced using the official implementation.

Method	Backbone	mA	F1	mADM	mAP	R-1
VAC [Guo19]†	ResNet-50	64.3±1.8	71.4±6.1	–	6.2±3.3	7.4±4.1
ALM [Tan19c]†	BN-Inception	72.0±2.3	78.0±2.7	–	11.4±4.2	13.9±9.0
SAL [Cao20]†	ResNet-50 + MLP	–	–	–	7.9±3.4	10.5±7.2
SB [Jia21b]†	ResNet-50	71.1±2.0	78.7±3.0	49.0±7.1	13.3±4.6	16.4±10.0
	ConvNeXt-Base	**73.2**±1.9	**82.1**±2.7	56.1±5.1	17.6±4.3	19.2±8.3
Proposed framework	ResNet-50	72.9±2.3	80.5±2.5	**56.4**±6.7	**20.2**±6.0	**21.3**±10.4
	ConvNeXt-Base	**75.1**±2.0	**83.9**±1.9	**61.2**±5.3	**27.2**±6.0	**28.0**±10.7

The proposed framework sets the state-of-the-art by a large margin. When comparing the methods that leverage identical backbones, improvements are observed regarding both tasks and all metrics. Especially concerning attribute-based person retrieval metrics, methods from the related literature achieve clearly worse results. For the experiments with the ResNet-50 and ConvNeXt-Base backbones, the gap to the best approach from the literature in mADM is 7.4 and 5.1 points, respectively. A further remarkable finding

is that the use of models with straightforward PAR architectures, such as SB [Jia21b] and the proposed framework, leads to superior performance compared to more complex models, *e.g.*, SAL [Cao20]. This demonstrates that intricate methods carry an increased risk of overfitting and overadapting to the training data and its characteristics, resulting in poor generalization.

Identical observations are made for the 4FCV evaluation protocol for which the results are provided in Table 8.9. In each split, a single sub-dataset of UPAR is utilized for training and three datasets are leveraged for evaluation.

Table 8.9: State-of-the-art UPAR 4FCV – Generalization results for the UPAR 4FCV protocol. The proposed framework outperforms the methods from related literature, particularly concerning the attribute-based person retrieval task. When two backbone models are stated, the first one represents the image, while the second one represents the text or attribute encoder. MLP refers to a small network of FC layers. Typically, three to four layers are used. **Red** and **blue** colors denote the best and second-best results, respectively. † Results were produced using the official implementation.

Method	Backbone	mA	F1	mADM	mAP	R-1
VAC [Guo19]†	ResNet-50	64.3±1.8	71.4±6.1	–	6.2±3.3	7.4±4.1
ALM [Tan19c]†	BN-Inception	66.3±2.0	71.0±5.4	–	5.5±3.0	7.3±4.0
SAL [Cao20]†	ResNet-50 + MLP	–	–	–	3.7±1.5	4.8±2.0
SB [Jia21b]†	ResNet-50	65.4±2.4	71.1±5.9	36.4±6.5	6.5±3.3	8.2±4.6
	ConvNeXt-Base	**70.1**±1.5	**77.2**±4.2	**46.4**±5.8	11.0±3.6	12.8±4.6
Proposed framework	ResNet-50	68.3±1.6	73.6±5.5	45.1±6.9	**11.7**±4.5	**13.1**±5.3
	ConvNeXt-Base	**70.9**±1.6	**79.7**±3.4	**51.7**±5.8	**19.0**±5.3	**19.9**±6.4

The proposed framework achieves the strongest results concerning both tasks. The second-best SB [Jia21b] method with the same backbone is outperformed by 8.7 and 5.3 points in mADM for ResNet-50 and ConvNeXt-Base, respectively.

In conclusion, the proposed framework demonstrates superiority over the state-of-the-art methods in related literature for both specialization and generalization scenarios. Notably, impressive outcomes are achieved for attribute-based person retrieval, but also w.r.t. PAR, even when design choices are made that are optimal for improving attribute-based person retrieval at the expense of PAR performance.

8.4 Summary

Finally, the results of this chapter and their implications for the real-world application of the proposed attribute-based person retrieval framework are discussed.

Combining the proposed approaches toward enhancing PAR (refer to Chapter 6) and attribute-based person retrieval (refer to Chapter 7) yields significant improvements. The results demonstrate that additional components added to the framework improve the outcomes, despite the similar objectives of individual optimizations. Whether investigating specialization or generalization, the combined methods lead to superior performance. The PARNorm module is the only exception, that is thus omitted from the final results. The analysis of the inference time of the proposed feature extraction model demonstrates little overhead compared to the baseline model. Using an *NVIDIA GeForce RTX 3090* GPU, more than 400 person images are processed per second with the *PyTorch* framework, which is expected to be sufficient for the use in real-world scenarios.

Representative search results were then examined to verify the effectiveness of the proposed methodology. The observations indicate that attribute recognition is significantly improved and that other components such as the HP, reliability calibration, and the DBD distance achieve their objectives. This applies to both the specialization and generalization scenarios. The study found that the suggested framework generates reliable retrieval rankings that are in line with human expectations. Moreover, it showcased the system's ability to retrieve matches in challenging scenarios where there is only a single matching sample for the query in the gallery dataset.

Regarding weaknesses and potential areas for improvement, it has been discovered that in rare cases, inaccurate hardness scores generated by the HP lead to some attributes being ignored during retrieval. Thus, some samples are ranked at early positions that do not match the query. Nevertheless, given that considerable enhancements are observed through the inclusion of the HP, this appears to be only a minor issue for a small percentage of queries.

The analysis also revealed that failure cases, where the proposed framework is unable to retrieve samples matching the query description, are caused by attributes with large intra-class variance, unbalanced attributes, and poor image quality due to low resolution, noise, or illumination. This demonstrates that although these issues have been addressed and significant improvements have been made, there is still room for further progress.

The comparison with current state-of-the-art methods showed that the proposed framework outperforms methods from the literature. Regardless of the selected backbone model, the proposed framework surpasses other methods in terms of attribute-based person retrieval on each of the specialization datasets, as well as on both UPAR evaluation protocols. In particular, the framework exhibits strong generalization performance, while more complex approaches from the literature struggle with overfitting. This demonstrates the adaptability and versatility of the approach for the use in a variety of real-world applications, regardless of the availability of training data from the target domain. Additionally, the investigation indicates that achieving top performance in the field of PAR does not require complex model architectures. The well-tuned baseline architecture introduced in this thesis is sufficient, even though it is not specifically optimized for the task.

9 Tracking System

To complete the concept of this thesis presented in Chapter 3, this chapter outlines a MTMCT system that incorporates the proposed framework for attribute-based person retrieval. Since the primary focus of this thesis is on the topics of PAR and person retrieval, only a concise overview of additional research conducted on MTMCT is presented. More details regarding the approaches are provided in the related publications [Köh20, Spe22a, Spe22c].

In general, MTMCT systems track the movements of multiple individuals captured by a network of cameras placed at different locations. The use of such a system in the context of this thesis provides several key advantages, including:

- The information regarding individuals is compressed, meaning entire movements of people can be retrieved based on attribute queries instead of single occurrences. This improves clarity, reduces manual effort, and enhances efficiency.
- The resulting tracks include an increased amount of information, thus enabling the use of video-based PAR methods, as detailed in Section 6.3. As a result, attribute predictions are more robust.

The processing pipeline of MTMCT approaches often follows the tracking-by-detection paradigm [Cia20]. Figure 9.2 visualizes this procedure based on the Weighted Distance Aggregation (WDA) tracker [Köh20]. Initially, individuals are detected and localized in the video frames generated by each camera. Subsequently, single-camera trackers are employed to connect the person detections showing the same individual over time in the same camera view. This stage commonly incorporates person re-identification models to generate feature vectors for each detected bounding box that encode the visual appearance

of individuals. This information enables the continuation of tracks after occlusions and the creation of multi-camera tracks by connecting single-camera tracks across different camera views. At this point, the PAR model proposed in Chapter 6 can be employed as an additional feature extractor. The video-based temporal pooling approach enables the aggregation of visual information over time, resulting in track-level predictions of soft biometric characteristics. Finally, the inter-camera association module, or the actual multi-camera tracker, combines the obtained single-camera tracks across multiple cameras. The resulting multi-camera tracks represent the entire trajectories of persons through the camera network, which are stored as the gallery in conjunction with the extracted semantic attributes. Utilizing the attribute-based person retrieval methodology presented in Chapter 7, this gallery database can be searched through with person descriptions as queries.

The remainder of this chapter provides a brief overview of the author's contributions to the research field of MTMCT of persons. Further works concerning vehicle tracking [Spe21b, Spe22b] are left out since the transfer to the person domain is not yet evaluated. First, the MTA dataset is presented to address the lack of large-scale MTMCT datasets [Köh20]. This is followed by the introduction of approaches for the offline [Köh20] and online [Spe22a] use cases presented in Chapter 1. Last, a specific system implementation for real-world deployment in real-time is described [Spe22c].

MTA dataset: The MTA dataset aims at solving the lack of proper datasets for MTMCT, as the existing datasets [Kuo10, Zha15] have a shortage of diverse person identities, capture a short video span, and lack diversity in environmental aspects like weather and illumination. Due to privacy concerns and tedious manual annotation processes, creating real-world datasets for MTMCT is a strenuous task. For instance, the DukeMTMC [Ris16] dataset was withdrawn due to inappropriate consent collection. To avoid such issues, the popular video game GTA V[1] was used to record the MTA dataset. It is argued that, contrary to direct image processing tasks, findings about MTMCT can

[1] https://www.rockstargames.com/de/gta-v

be transferred from synthetic to real-world data with a significantly smaller domain gap.

The dataset illustrates an urban scene recorded by six cameras, as displayed in Figure 9.1. The dataset is unique compared to existing MTMCT datasets, since it contains both overlapping as well as non-overlapping camera views. Additionally, it exhibits an extensive range of characteristics, such as daytime and nighttime, indoor and outdoor areas, varying levels of crowds, and fluctuating weather conditions.

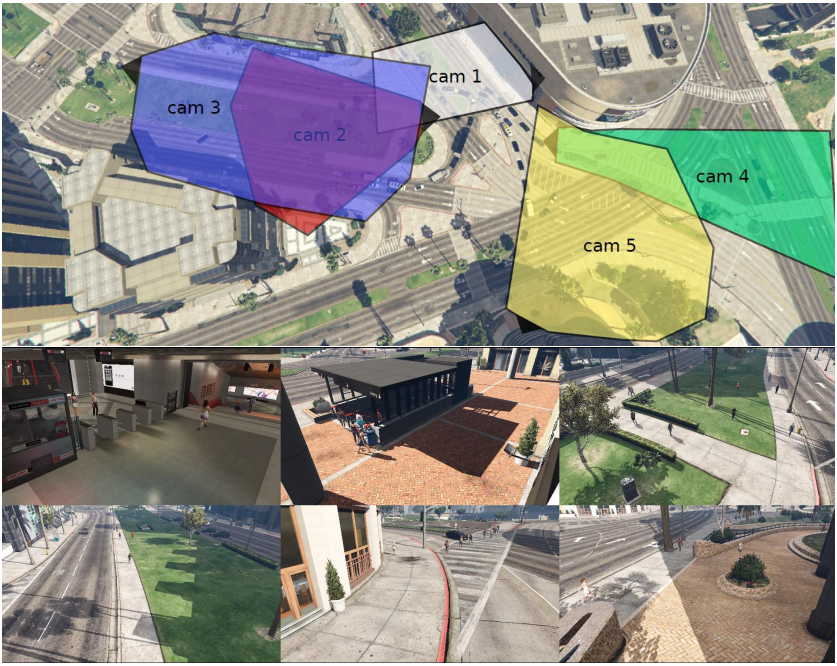


Figure 9.1: MTA dataset – The upper figure depicts the fields of view of five out of six cameras that were employed in gathering the MTA dataset via the GTA V video game. The remaining camera is positioned indoors in a subway station near camera 1. The camera perspectives are visualized in the bottom part.

In total, the datasets comprises 10 hours of video footage depicting over 2,800 distinct individuals, featuring highly accurate annotations that include person IDs, 2D and 3D person positions, as well as labels for 22 human body joints. Therefore, the MTA dataset allows for comprehensive evaluations of MTMCT techniques under real-world conditions, including varying weather conditions. The dataset served as the basis for developing the offline and online tracking approaches detailed in the following.

Offline tracking: Offline tracking addresses the forensic use case, meaning that MTMCT and investigations are conducted after all necessary data has been collected. As a post-processing step, offline trackers can solve the task without a real-time processing requirement, and all data is readily available. Consequently, offline tracking is generally considered easier and achieves better accuracy than online algorithms. Thus, most research in the related literature concentrates on this type of processing [Ris18, He20, Ye21].

The WDA offline tracker [Köh20] was developed as part of this thesis. An overview is provided in Figure 9.2. Given the multitude of established methods available for the person detection [Ren16, Fen21, Zha21a, Wan22a, Joc23] and single-camera tracking [Ber19, Zho20, Aha22, Du23] stages, the WDA tracker focuses on the inter-camera association part of the processing pipeline. To associate single-camera tracks across the network of cameras, agglomerative hierarchical clustering is utilized. Basically, this method connects the two most similar tracks iteratively until a stopping criterion is met. The WDA tracker employs five different criteria to assess the similarity, which leverage prior knowledge about the scene. Without this information, accomplishing strong MTMCT accuracy becomes challenging due to the massive complexity of the task. The clustering process stops when the similarity criteria between the closest tracks drops below a threshold. The five criteria that contribute to the similarity measure are the following ones:

- **Visual similarity:** The visual similarity of peoples' overall appearance is key for rediscovering persons who exit one camera's view and enter another. The person re-identification model extracts feature representations to facilitate this process. Moreover, semantic

attributes predicted by a PAR may serve as complementary features to discern the visual similarity between tracks.

- **Single-camera time constraint:** Individuals cannot appear within multiple tracks on the same camera simultaneously. This temporal constraint prevents the linking of such single-camera tracks.
- **Multi-camera time constraint:** Similar to the single-camera time constraint, it is not possible for one person to appear simultaneously in multiple cameras with non-overlapping fields of view. Thus, such combinations are avoided.
- **Homography matching:** For the overlapping camera views, individuals' positions within the tracks are transformed into the overlapping camera views and then compared with the active tracks in these cameras at corresponding times. The level of similarity is assessed by measuring the agreement between the transformed track positions and track detections in the overlapping camera.
- **Movement prediction:** Last, predictions of persons further movements are made once a single-camera track has ended. The similarity of tracks qualifying as potential successors is increased when initiated near the spatio-temporal predictions.

Each of the criteria is formulated as a distance metric and the weighted sum of these distances is leveraged to determine the similarity between tracks. The experimental results show a notable improvement in the MTMCT evaluation metric Identity F1 (IDF1) [Ris16] on the MTA dataset, with a rise from 17.3% to 30.1%. The IDF1 score quantifies the tracking accuracy by computing the ratio of successfully identified detections to the average number of ground truth and produced detections. These findings underscore the importance of incorporating knowledge regarding camera placement to obtain robust MTMCT accuracy. Furthermore, the modular WDA framework provides a solid foundation to enhance MTMCT by including additional similarity measures and constraints.

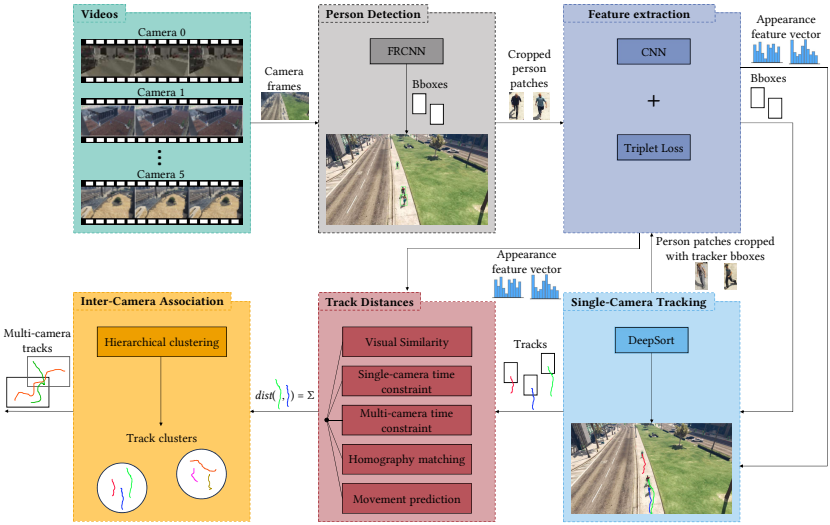


Figure 9.2: WDA tracker – Overview of the functional principle of the WDA tracker. It follows the tracking-by-detection paradigm. The inter-camera association relies on the weighted aggregation of multiple distances. Changed representation after [Köh20].

Online tracking: In contrast to offline approaches, online tracking must address the MTMCT task on a frame-by-frame basis, *i.e.*, tracks must be associated across cameras immediately after their first appearance. Instances that have previously been observed in another camera must be re-identified as quickly as possible without having much information about the new track. This enables real-time analytics and attribute-based person searches of tracks in live situations. For instance, criminals matching a certain person description can be found and tracked in the live feeds to arrest them. Nonetheless, global clustering approaches are infeasible for online scenarios since the information is incrementally available. Therefore, the WDA tracker's inter-camera association component is replaced for online processing [Spe22a]. Compared to offline MTMCT, there are limited studies in the literature that focus on

the online use case [Zha19a, Gai21, Ric23]. Additionally, these studies either ignore vital scene information and concentrate solely on visual appearance features [Gai21] or tackle the specialized task where all cameras overlap [Zha19a, Ric23]. The proposed online MTMCT approach can handle both overlapping and non-overlapping cameras and improves tracking accuracy through a scene model. It follows a similar procedure to the WDA tracker for person detection, single-camera tracking, and similarity computation, but differs in the inter-camera association algorithm. A probability-based association procedure is introduced, replacing the clustering method. Connecting a new track to a multi-camera track displaying the same person as early as possible is preferred. However, when someone enters a camera view, the person is only seen from a single viewing angle, resulting in a weak appearance representation. Thus, directly assigning the track to a multi-camera track may increase the chance of erroneous associations. On the other hand, delaying the decision for too long leads to offline tracking and hinders the goal of real-time person search.

To achieve a suitable tradeoff and address this issue objectively, it is suggested making the assignment once enough information is available to make an informed decision. This involves connecting a new track within a camera to a potential set of predecessors instead of assigning it directly to a certain multi-camera track. At each subsequent time step, the similarity between the new track and the potential predecessors is recalculated to refine the set. For instance, unlikely tracks and tracks that have since been matched with another track that violates one of the spatio-temporal constraints are removed. Once only one multi-camera track remains or one track is much more likely to be the predecessor than the others in the set, the assignment is carried out. Experimental results confirm that decisions are generally made within the first few frames, as the scene model successfully eliminates impossible combinations and limits the feasible predecessors. Furthermore, experimental results prove the benefits concerning difficult assignment decisions.

The proposed methodology surpasses Gaikwad et al. [Gai21]'s approach by 4.3 percentage points in IDF1 on the MTA dataset. In terms of the performance difference between offline and online trackers, the proposed online approach

scored 2.1 points lower in IDF1 compared to the offline WDA tracker, which appears acceptable considering the advantages for the online use case. The approach demonstrated its capability for real-time processing since it achieved a processing speed of over 37 Frame Per Second (FPS) in the experiments processing six cameras simultaneously.

Real-world person retrieval system: Finally, covering the processing pipeline introduced in Chapter 3, the following briefly outlines the implementation of a real-world person retrieval system [Spe22c]. Deploying these systems in real-world scenarios presents additional challenges and requirements compared to research. Specifically, when used in online settings, these systems need to satisfy real-time processing. Achieving this goal can be challenging due to budget constraints that necessitate utilizing affordable hardware. Additionally, it is essential to ensure high levels of flexibility and scalability that enable seamless integration of additional camera streams or processing hardware. Moreover, compliance with data protection regulations is imperative.

An overview of the system implementation is given in Figure 9.3. The system utilizes Docker[1], a container virtualization software, for its flexible scalability and multi-server deployment capabilities. The system comprises distinct modules that intercommunicate with each other. Each color depicted in the figure characterizes a separate Docker image, and each box represents a separate Docker container instance. The system can be readily expanded by connecting additional servers and managing camera streams through starting or stopping Docker containers. It should be noted that the inter-camera association of tracks was added to the system after the publication of this paper and is therefore not included in the overview.

The single-camera processing module comprises components for person detection, single-camera tracking, and feature extraction. In the context of this thesis, feature extraction includes creating global appearance feature vectors

[1] https://www.docker.com/

utilizing person re-identification techniques and extracting the semantic attributes for individuals via the video-based PAR model introduced in Chapter 6.

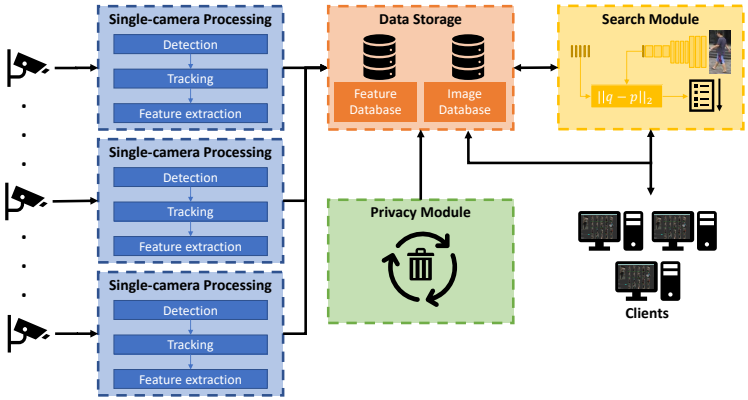


Figure 9.3: Real-world person retrieval system – The implementation of a real-world person retrieval system is depicted. It consists of multiple stages. First, persons are tracked and features, such as the individuals' semantic attributes, are extracted. The collected data stored in a gallery database can then be queried using clients and a separate search module. Furthermore, a privacy modules ensures compliance with legal regulations. Figure source: [Spe22c].

The data storage module maintains data for the tracks, including metadata about time and place of occurrence, re-identification features along with semantic attributes, and image data for visualization purposes. This module serves as the gallery database for subsequent attribute-based person searches.

A privacy module is utilized to adhere to the European Union's General Data Protection Regulation (GDPR) through the strict enforcement of a maximum storage period. In addition, if the privacy module identifies any problems or is disabled by a user, the entire system is immediately terminated, and system operators are notified.

The search module responds to client search requests and performs person retrievals as explained in Chapter 7. This module offers two types of retrieval: the typical person search and a watch list search. In a typical person search, a ranked list of tracks in the database is generated based on the similarity of the query to the stored tracks. In contrast, solely new entries in the database are taken into account during a watch list search. Clients receive a notification along with its associated metadata if the distance between the query and a track falls below a particular threshold.

An extensive analysis of the system and assessment of various processing components and configurations verified the proposed retrieval system's ability to operate in real-time.

Summary: In summary, this chapter outlines the necessary components for a complete real-world attribute-based person retrieval system. The chapter introduces the MTMCT task and the MTA dataset, which addresses the shortage of suitable research datasets for MTMCT. Furthermore, this chapter provides a brief description of two multi-camera tracking methods, each of which addresses one of the two use cases motivated in Chapter 1, namely retrograde and online investigations. Last, a concrete implementation of such a system is presented, including a description of the integration of the attribute-based person retrieval framework proposed in Chapters 6 and 7.

10 Conclusion and Outlook

After the evaluation of the framework for attribute-based person retrieval proposed in this thesis, the conclusion and outlook chapter summarizes and reflects the main findings and outcomes in Section 10.1. Following this, Section 10.2 provides an outlook on potential future developments.

10.1 Conclusion

In this thesis, a novel deep learning-based framework for attribute-based person retrieval is proposed. This framework enables efficient search in mass data collected by multi-camera networks. As the base approach for solving this task, the use of PAR methods to extract soft biometric characteristics of persons is chosen. First, various design characteristics of PAR approaches are systematically studied for their impact on attribute-based person retrieval. Particularly, using SWA, the focal loss function, and reducing the batch size for small datasets are found to significantly improve the results of attribute-based person retrieval. However, optimizing a PAR model for attribute-based person retrieval may lead to suboptimal recognition of individual attributes. To counteract this, the PARNorm module proposes normalizing the model's output logits in two distinct ways. First, it normalizes them attribute by attribute along the batch dimension to increase the positive recall of attribute recognition. Then, since this degrades instance-based PAR metrics and person retrieval, instance-wise normalization is performed along the attribute dimension to avoid overestimating the number of present attributes. The experimental results illustrate that the module is effective in improving the recognition of individual soft biometric characteristics, while also enhancing attribute-based person retrieval. Furthermore, camera networks typically offer video

feeds rather than single images, providing more comprehensive information about a person's appearance as the individual is captured multiple times from various perspectives. This thesis argues that straightforward temporal average pooling is adequate for video-based PAR since relevant soft biometrics are not motion dependent. Experimental validation confirms the hypothesis. The proposed method of pooling global backbone features over time outperforms temporal attention models, 3D architectures, and transformer-based methods, while offering favorable inference time.

To further enhance the accuracy and reliability of attribute-based person retrieval rank lists, an independent HP is proposed. The HP is implemented as a distinct model branch and is trained to predict the difficulty of recognizing semantic attributes in the input image. It is demonstrated that the quality of retrieval rank lists is significantly improved by using this complementary information to focus on reliably recognized attributes during the generation of retrieval results. The analysis of difficulty predictions and the comparison with self-referential hardness prediction methods show the superiority of the proposed independent HP, as it more accurately identifies challenging factors in the input image that raise the chance of failure in attribute recognition. Besides, three additional enhancements to the retrieval process are investigated. Reliability calibration aims to align the confidence scores of attribute presence from the PAR model with the empirical probabilities of attribute presence. This eliminates over- or underconfidence in predictions caused by imbalanced attribute distributions and limited intra-class diversity in the training set. Applying such methods before computing the retrieval distance proves beneficial. However, the effectiveness of this approach relies on the availability of suitable validation data. Additionally, a weighting mechanism is proposed which balances the influence of attributes on the retrieval results by compensating for the differing expected errors produced by the attributes. Similar to reliability calibration, noteworthy improvements are observed. Last, a distance measure is proposed that takes into account the actual output distributions of the PAR classifier for the presence and absence of attributes. For this, logistic distributions are fitted to the classifier's output distributions. The resulting cumulative density and survival functions are then utilized to evaluate

the probability that the attribute predictions match the query. It is noteworthy that compared to the use of annotated validation data, the technique of estimating distribution parameters by generating pseudo-labels for data from the target domain has shown superior performance. This can be attributed to the existence of a domain gap between the validation and testing data.

The lack of uniform attribute annotations across datasets prevents generalization experiments with training and evaluation data from different domains. To evaluate the generalizability of PAR and attribute-based person retrieval methods, four research datasets are harmonized to form the UPAR dataset by contributing more than 3.3 million new binary soft biometric annotations and two evaluation protocols. Furthermore, this thesis claims that current metrics for evaluating attribute-based person retrieval omit important information as they do not take into account the degree of match between the query and retrieved gallery samples. The novel mADM metric resembles mAP, but accounts for the degree of correspondence between gallery samples and query attributes. It also introduces a normalization procedure to reduce the dependence of the resulting scores on the specific gallery, thus improving the comparability of the obtained results across datasets.

To ensure accurate retrieval, the proposed methods are combined into a unified framework. With the exception of the PARNorm module, which solely enhances PAR when used together with optimized design decisions, the combination proves to be advantageous. Furthermore, experiments exploring inference times demonstrate the efficient computation of the proposed PAR approach with little overhead over the baseline. The proposed framework surpasses representative works from the literature concerning both PAR and attribute-based person retrieval. The results demonstrate that this applies not only to the individual research datasets but also to the generalization experiments with the UPAR dataset.

Finally, an entire system for attribute-based person retrieval is outlined. The system incorporates the proposed framework and covers the full process from video feeds of cameras in the network to retrieval rank lists displayed to the system operators.

10.2 Outlook

Despite achieving remarkable outcomes regarding attribute-based person retrieval, additional enhancements and extensions could lead to further improvements and increase the practical applicability of the framework.

The analysis revealed that large and diverse training datasets are key to improving the generalization capabilities of attribute-based person retrieval methods, which is crucial for robust results in real-world scenarios. Consequently, extending the existing UPAR dataset and incorporating new data sources with diverse characteristics is a promising research direction. For instance, the dataset contains indoor and outdoor imagery taken during both daytime and nighttime. However, it currently lacks imagery captured in different weather conditions, such as rain or snow. In addition, utilizing drone-based cameras to capture mass events is on the rise, so the inclusion of drone-based camera footage may broaden the applicability in different scenarios. Furthermore, annotating additional soft biometric characteristics to the existing UPAR data would enable more detailed searches. Particular attention should be paid to differentiating between multiple types of clothing patterns, as the evaluation revealed problems when searching for persons wearing such garments. In addition, the MARS [Zhe16, Che19] dataset represents the only video-based surveillance dataset for PAR. Expanding the concept of the UPAR dataset to incorporate a video-based version is considered advantageous in order to allow generalization experiments and to validate results on different datasets. For instance, the MEVID dataset [Dav23] is an excellent choice since it features a variety of scenarios that accurately reflect real-world situations.

Another finding suggests that the recognition of individual attributes, as reflected by the label-based mA metric, has potential for enhancement. Qualitative evaluation results also indicate possible benefits for attribute-based person retrieval. However, it was observed that employing techniques that increase the mA can negatively impact the retrieval process. Considering the straightforward architecture of the PAR model in the proposed framework, it is worth exploring the integration of additional modules, such as attention

mechanisms or spatial projection components to improve localization of attributes and account for varying spatial extents of attributes. According to initial studies by the author of this thesis [Spe23a], positive effects are suggested on both PAR and attribute-based person retrieval tasks. Nonetheless, it is essential to consider potential implications, such as increased model sizes and inference times, to ensure compatibility with real-world deployment requirements.

Continuous improvement of the attribute-based person retrieval system is essential during practical application. Even with large-scale training datasets, there may not be enough diversity to guarantee models that can generalize effectively to all scenarios. Furthermore, specific characteristics or issues may arise only in the target domain. At present, the person retrieval system lacks a continuous learning component. Therefore, investigating, developing, and integrating continuous learning algorithms [Lan22, Wan23b] for PAR and attribute-based person retrieval into the system can enhance the system's usefulness and robustness iteratively during operation.

Bibliography

[Abu16] Abu-El-Haija, Sami; Kothari, Nisarg; Lee, Joonseok; Natsev, Paul; Toderici, George; Varadarajan, Balakrishnan and Vijayanarasimhan, Sudheendra: "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016) (cit. on p. 128).

[Aga22] Agarwal, Chirag; D'souza, Daniel and Hooker, Sara: "Estimating Example Difficulty using Variance of Gradients". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (New Orleans, LA, USA). IEEE Computer Society, 6/18/2022 - 6/24/2022, pp. 10358–10368. doi: 10.1109/CVPR52688.2022.01012 (cit. on p. 42).

[Agg20] Aggarwal, Surbhi; Babu, R. Venkatesh and Chakraborty, Anirban: "Text-based Person Search via Attribute-aided Matching". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020. doi: 10.1109/WACV45572.2020.9093640 (cit. on p. 28).

[Aha22] Aharon, Nir; Orfaig, Roy and Bobrovsky, Ben-Zion: "BoT-SORT: Robust Associations Multi-Pedestrian Tracking". In: *arXiv preprint arXiv:2206.14651* (2022) (cit. on p. 200).

[Akt23] Aktuell, S.W.R.: Videoüberwachung in Mannheim mit positivem Effekt. Copyright: SWR/Südwestrundfunk - Anstalt des Öffentlichen Rechts. SWR Aktuell. 2023. url: https://www.swr.de/swraktuell/baden-wuerttemberg/mannheim/

videoueberwachung-mannheim-erfolgsgeschichte-100.html (visited on 11/01/2023) (cit. on p. 2).

[Arp22] Arpit, Devansh; Wang, Huan; Zhou, Yingbo and Xiong, Caiming: "Ensemble of Averages: Improving Model Selection and Boosting Performance in Domain Generalization". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh; Alekh Agarwal; Danielle Belgrave and Kyunghyun Cho. 2022. URL: https://openreview.net/forum?id=peZSbfNnBp4 (cit. on pp. 90, 91, 113).

[Arr23] Arriaga, Octavio; Palacio, Sebastian and Valdenegro-Toro, Matias: "Difficulty Estimation with Action Scores for Computer Vision Tasks". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Vancouver, BC, Canada). IEEE, 6/17/2023 - 6/24/2023, pp. 245–253. DOI: 10.1109/CVPRW59228.2023.00030 (cit. on p. 42).

[Ba16] Ba, Jimmy Lei; Kiros, Jamie Ryan and Hinton, Geoffrey: Layer Normalization. 21.07.2016. URL: https://arxiv.org/pdf/1607.06450 (cit. on pp. 40, 41, 117).

[BBC13] BBC News: "Boston Marathon bombings: How notorious bombers got caught". In: *BBC News* (Apr. 18, 2013). URL: https://www.bbc.com/news/magazine-22191033 (visited on 11/22/2022) (cit. on p. 2).

[Ber19] Bergmann, Philipp; Meinhardt, Tim and Leal-Taixe, Laura: "Tracking without bells and whistles". In: *arXiv preprint arXiv:1903.05625* (2019) (cit. on p. 200).

[Bla11] Blanchard, Gilles; Lee, Gyemin and Scott, Clayton: "Generalizing from Several Related Classification Tasks to a New Unlabeled Sample". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor; R. Zemel; P. Bartlett; F. Pereira and K.Q. Weinberger. Vol. 24. Curran Associates, Inc,

2011. URL: https://proceedings.neurips.cc/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf (cit. on p. 72).

[Bou11] Bourdev, Lubomir; Maji, Subhransu and Malik, Jitendra: "Describing people: A poselet-based approach to attribute classification". In: *2011 International Conference on Computer Vision (ICCV 2011). Barcelona, Spain, 6 - 13 November 2011.* 2011 IEEE International Conference on Computer Vision (ICCV) (Barcelona, Spain). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2011, pp. 1543–1550. DOI: 10.1109/ICCV.2011.6126413 (cit. on p. 31).

[Bov05] Bovik, Alan C.: Handbook of Image and Video Processing. eng. 2nd ed. Communications, Networking and Multimedia Ser. Bovik, Alan C. (VerfasserIn). Burlington: Elsevier Science & Technology, 2005. 1429 pp. URL: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=328547 (cit. on p. 8).

[Cao20] Cao, Yu-Tong; Wang, Jingya and Tao, Dacheng: "Symbiotic Adversarial Learning for Attribute-Based Person Search". In: *Computer Vision – ECCV 2020. 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV.* Ed. by Vedaldi, Andrea; Bischof, Horst; Brox, Thomas and Frahm, Jan-Michael. 1st ed. 2020. Vol. 12359. Springer eBook Collection 12359. Cham: Springer International Publishing and Imprint Springer, 2020, pp. 230–247. DOI: 10.1007/978-3-030-58568-6_14 (cit. on pp. 37, 38, 46, 191–193).

[Car19] Carreira, Joao; Noland, Eric; Hillier, Chloe and Zisserman, Andrew: "A short note on the kinetics-700 human action dataset". In: *arXiv preprint arXiv:1907.06987* (2019) (cit. on p. 128).

[Cha21] Cha, Junbum; Chun, Sanghyuk; Lee, Kyungjae; Cho, Han-Cheol; Park, Seunghyun; Lee, Yunsung and Park, Sungrae: "SWAD: Domain Generalization by Seeking Flat Minima". In: *Advances in Neural Information Processing Systems (NeurIPS).* 2021 (cit. on p. 92).

[Che18] Chen, Tianlang; Xu, Chenliang and Luo, Jiebo: "Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (Lake Tahoe, NV). IEEE, 2018, pp. 1879–1887. DOI: 10.1109/WACV.2018.00208 (cit. on p. 27).

[Che19] Chen, Zhiyuan; Li, Annan and Wang, Yunhong: "A Temporal Attentive Approach for Video-Based Pedestrian Attribute Recognition". In: *Pattern recognition and computer vision. Second Chinese Conference, PRCV 2019, Xi'an, China, November 8-11, 2019 : proceedings*. Ed. by Lin, Zhouchen; Wang, Liang; Yang, Jian; Shi, Guangming; Tan, Tieniu; Zheng, Nanning; Chen, Xilin and Zhang, Yanning. Vol. 11858. Lecture notes in computer science 11858. Cham: Springer, 2019, pp. 209–220. DOI: 10.1007/978-3-030-31723-2_18 (cit. on pp. 35, 37, 39, 54, 64, 125, 127, 129, 131, 132, 169, 210).

[Che21] Chen, Yucheng; Huang, Rui; Chang, Hong; Tan, Chuanqi; Xue, Tao and Ma, Bingpeng: "Cross-Modal Knowledge Adaptation for Language-Based Person Search". eng. In: *IEEE Transactions on Image Processing* 30 (2021). Journal Article, pp. 4057–4069. DOI: 10.1109/TIP.2021.3068825. eprint: 33788687 (cit. on p. 28).

[Che22a] Chen, Yuhao; Zhang, Guoqing; Lu, Yujiang; Wang, Zhenxing and Zheng, Yuhui: "TIPCB: A simple but effective part-based convolutional baseline for text-based person search". In: *Neurocomputing* 494 (2022). PII: S0925231222004726, pp. 171–181. DOI: 10.1016/j.neucom.2022.04.081 (cit. on p. 28).

[Che22b] Cheng, Xinhua; Jia, Mengxi; Wang, Qian and Zhang, Jian: "A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.10 (2022), pp. 6994–7004. DOI: 10.1109/TCSVT.2022.3178144 (cit. on pp. 35, 190).

[Cia20] Ciaparrone, Gioele; Luque Sánchez, Francisco; Tabik, Siham; Troiano, Luigi; Tagliaferri, Roberto and Herrera, Francisco: "Deep learning in video multi-object tracking: A survey". In: *Neurocomputing* 381 (2020). PII: S0925231219315966, pp. 61–88. doi: 10.1016/j.neucom.2019.11.023 (cit. on p. 197).

[Cor01] Cordis: Safer crowds in mass gatherings. en. Publication Office/CORDIS. 2023-11-01. url: https://cordis.europa.eu/article/id/410936-safer-crowds-in-mass-gatherings (visited on 11/01/2023) (cit. on p. 2).

[Cor23] Cormier, Mickael; Specker, Andreas; Jacques, Julio C. S.; Florin, Lucas; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge: Pedestrian Attribute Recognition and Attribute-based Person Retrieval - Dataset, Design, and Results". In: *2023 IEEE Winter Conference on Applications of Computer Vision workshops. WACVW 2023 : proceedings : 3-7 January 2023, Waikoloa, Hawaii.* 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers and Computer Vision Foundation. Piscataway, NJ: IEEE, 2023, pp. 166–175. doi: 10.1109/WACVW58289.2023.00022 (cit. on pp. 14, 15, 48, 49, 59, 84, 116).

[Daf16] Daftry, Shreyansh; Zeng, Sam; Bagnell, J. Andrew and Hebert, Martial: "Introspective perception: Learning to predict failures in vision systems". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 1743–1750 (cit. on p. 41).

[Dan11] Dantcheva, Antitza; Velardo, Carmelo; D'Angelo, Angela and Dugelay, Jean-Luc: "Bag of soft biometrics for person identification". In: *Multimedia Tools and Applications* 51.2 (2011). PII: 635, pp. 739–777. doi: 10.1007/s11042-010-0635-7 (cit. on p. 18).

[Dan16] Dantcheva, Antitza; Elia, Petros and Ross, Arun A.: "What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics". In: *IEEE Transactions on Information Forensics and Security* 11.3 (2016), pp. 441–467. doi: 10.1109/TIFS.2015.2480381 (cit. on pp. 4, 9, 18–21, 60).

[Dav23] Davila, Daniel et al.: "MEVID: Multi-view Extended Videos with Identities for Video Person Re-Identification". In: *2023 IEEE Winter Conference on Applications of Computer Vision. 3-7 January 2023, Waikoloa, Hawaii : proceedings*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (Waikoloa, HI, USA). Ed. by Berg, Tamara. IEEE Computer Society and Computer Vision Foundation. Piscataway, NJ: IEEE, 2023, pp. 1634–1643. doi: 10.1109/WACV56688.2023.00168 (cit. on p. 210).

[Den14] Deng, Yubin; Luo, Ping; Loy, Chen Change and Tang, Xiaoou: "Pedestrian Attribute Recognition At Far Distance". In: *Proceedings of the 22nd ACM international conference on Multimedia*. the ACM International Conference (Orlando, Florida, USA). Ed. by Hua, Kien A. New York, NY: ACM, 2014, pp. 789–792. doi: 10.1145/2647868.2654966 (cit. on pp. 7–9, 11, 12, 30, 48, 53, 54, 60, 65, 184, 189).

[Den19] Deng, Jiankang; Guo, Jia; Xue, Niannan and Zafeiriou, Stefanos: "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2019 : 16-20 June 2019, Long Beach, California : proceedings*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA). Ed. by Davis, Larry. Institute of Electrical and Electronics Engineers and Computer Vision Foundation. Piscataway, NJ: IEEE, 2019, pp. 4685–4694. doi: 10.1109/CVPR.2019.00482 (cit. on p. 38).

[Dev19] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton and Toutanova, Kristina: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North*. Proceedings of the 2019 Conference of the North (Minneapolis, Minnesota). Ed. by Burstein, Jill; Doran, Christy and Solorio, Thamar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/n19-1423 (cit. on pp. 28, 35).

[Dib16] Diba, Ali; Pazandeh, Ali Mohammad; Pirsiavash, Hamed and van Gool, Luc: "DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns". In: *29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA). Ed. by Bajcsy, Růžena; Li, Fei-Fei and Tuytelaars, Tinne. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2016, pp. 3557–3565. doi: 10.1109/CVPR.2016.387 (cit. on p. 32).

[Din21] Ding, Zefeng; Ding, Changxing; Shao, Zhiyin and Tao, Dacheng: Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. A new database for text-to-image ReID is provided. Code will be released. 2021. url: https://arxiv.org/pdf/2107.12666 (cit. on pp. 27, 28).

[Don19] Dong, Qi; Zhu, Xiatian and Gong, Shaogang: "Person Search by Text Attribute Query As Zero-Shot Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019). IEEE, 2019. doi: 10.1109/ICCV.2019.00375 (cit. on pp. 37, 191).

[Dos21] Dosovitskiy, Alexey et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, 2021. doi: 10.48550/arXiv.2010.11929 (cit. on p. 86).

[dpa20] DPA: "Polizei Mannheim: Smarte Videoüberwachung hilft – aber kein Allheilmittel". In: *heise online* (Aug. 24, 2020). URL: https://www.heise.de/news/Polizei-Mannheim-Smarte-Videoueberwachung-hilft-aber-kein-Allheilmittel-4876470.html (visited on 11/24/2022) (cit. on p. 2).

[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* (2023) (cit. on p. 200).

[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: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 430–446. URL: http://openaccess.thecvf.com/content_ECCV_2018/papers/Matteo_Fabbri_Learning_to_Detect_ECCV_2018_paper.pdf (cit. on p. 60).

[Fan23] Fan, Xinwen; Zhang, Yukang; Lu, Yang and Wang, Hanzi: "PARFormer: Transformer-based Multi-Task Network for Pedestrian Attribute Recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023), p. 1. DOI: 10.1109/TCSVT.2023.3285411 (cit. on pp. 35, 39, 190).

[Fen17] Fennelly, Lawrence J. and Perry, Marianna A.: "Crime and Crime Prevention Techniques". In: *Physical Security: 150 Things You Should Know*. Elsevier, 2017, pp. 97–113. DOI: 10.1016/B978-0-12-809487-7.00003-6 (cit. on pp. 2, 20).

[Fen21] Feng, Chengjian; Zhong, Yujie; Gao, Yu; Scott, Matthew R. and Huang, Weilin: "Tood: Task-aligned one-stage object detection". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society %–VarifocalNet. 2021, pp. 3490–3499 (cit. on p. 200).

[Fer14] Feris, Rogerio; Bobbitt, Russel; Brown, Lisa and Pankanti, Sharath: "Attribute-based People Search". In: *Proceedings of International Conference on Multimedia Retrieval*. ICMR '14: International Conference on Multimedia Retrieval (Glasgow United Kingdom). Ed. by Jose, Joemon. ACM Digital Library. Association for Computing Machinery-Digital Library and ACM Special Interest Group on Multimedia. New York, NY: ACM, 2014, pp. 153–160. doi: 10.1145/2578726.2578732 (cit. on p. 3).

[Fli86] Flin, Rhona H. and Shepherd, John W.: "Tall stories: Eyewitnesses' ability to estimate height and weight characteristics". In: *Human Learning: Journal of Practical Research & Applications* (1986) (cit. on pp. 18, 24, 60, 154).

[Flo21] Florin, Lucas; Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "Hardness Prediction for More Reliable Attribute-based Person Re-identification". In: *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 418–424. doi: 10.1109/MIPR51284.2021.00077 (cit. on pp. 14, 50, 135).

[Gad23] Gadepally, Krishna Chaitanya; Bhusan Dhal, Sambandh; Kalafatis, Stavros and Nowka, Kevin J.: "Realistic Predictors for Regression and Semantic Segmentation". In: *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*. 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA) (Orlando, FL, USA). IEEE, 5/23/2023 - 5/25/2023, pp. 153–155. doi: 10.1109/SERA57763.2023.10197824 (cit. on pp. 42, 135).

[Gai21] Gaikwad, Bipin and Karmakar, Abhijit: "Smart surveillance system for real-time multi-person multi-camera tracking at the edge". In: *Journal of Real-Time Image Processing* (2021). 02. doi: 10.1007/s11554-020-01066-8 (cit. on p. 203).

[Gal16] Gal, Yarin and Ghahramani, Zoubin: "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. JMLR.org, 2016, pp. 1050–1059 (cit. on p. 42).

[Gal21] Galiyawala, Hiren and Raval, Mehul S.: "Person retrieval in surveillance using textual query: a review". In: *Multimedia Tools and Applications* 80.18 (2021). PII: 10983, pp. 27343–27383. doi: 10.1007/s11042-021-10983-0 (cit. on pp. 3, 4, 24).

[Gao21] Gao, Shang-Hua; Cheng, Ming-Ming; Zhao, Kai; Zhang, Xin-Yu; Yang, Ming-Hsuan and Torr, Philip: "Res2Net: A New Multi-Scale Backbone Architecture". eng. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2021). Journal Article Research Support, Non-U.S. Gov't Journal Article Research Support, Non-U.S. Gov't, pp. 652–662. doi: 10.1109/TPAMI.2019.2938758. eprint: 31484108 (cit. on pp. 77, 86).

[Gaw21] Gawlikowski, Jakob et al.: A Survey of Uncertainty in Deep Neural Networks. 7.07.2021. url: https://arxiv.org/pdf/2107.03342.pdf (cit. on p. 42).

[Gen21] Geng, Chuanxing; Huang, Sheng-Jun and Chen, Songcan: "Recent Advances in Open Set Recognition: A Survey". eng. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021). Journal Article Research Support, Non-U.S. Gov't Journal Article Research Support, Non-U.S. Gov't, pp. 3614–3631. doi: 10.1109/TPAMI.2020.2981604. eprint: 32191881 (cit. on p. 42).

[Gir15] Girshick, Ross: "Fast R-CNN". In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015). IEEE, 2015. doi: 10.1109/ICCV.2015.169 (cit. on p. 31).

[Gki15] Gkioxari, Georgia; Girshick, Ross and Malik, Jitendra: "Actions and Attributes from Wholes and Parts". In: *2015 IEEE International Conference on Computer Vision. 11-18 December 2015,*

Santiago, Chile : proceedings. 2015 IEEE International Conference on Computer Vision (ICCV) (Santiago, Chile). Ed. by Bajcsy, Ruzena; Hager, Greg and Ma, Yi. IEEE International Conference on Computer Vision and Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2015, pp. 2470–2478. doi: 10.1109/ICCV.2015.284 (cit. on pp. 31, 39, 48, 83).

[Gle22] Glenn Jocher et al.: ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference. feb, 2022. doi: 10.5281/zenodo.6222936 (cit. on p. 91).

[Gol22] Golda, Thomas; Guaia, Deborah and Wagner-Hartl, Verena: "Perception of Risks and Usefulness of Smart Video Surveillance Systems". In: *Applied Sciences* 12.20 (2022). PII: app122010435, p. 10435. doi: 10.3390/app122010435 (cit. on p. 2).

[Gol23] 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. 2., vollständig überarbeitete und werweiterte Auflage. Wiesbaden: Springer Gabler, 2023, pp. 1487–1507. doi: 10.1007/978-3-658-34388-0_87 (cit. on p. 2).

[Gra07] Gray, Douglas; Brennan, Shane and Tao, Hai: "Evaluating appearance models for recognition, reacquisition, and tracking". In: *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*. Vol. 3. 2007, pp. 1–7 (cit. on pp. 53, 54).

[Guo16] Guo, Yandong; Zhang, Lei; Hu, Yuxiao; He, Xiaodong and Gao, Jianfeng: "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition". In: *Computer vision - ECCV 2016. 14th European conference, Amsterdam, The Netherlands, October 11-14, 2016 : proceedings*. Ed. by Leibe, Bastian; Matas, Jiri; Sebe, Nicu and Welling, Max. Vol. 9907. Lecture notes in computer science 9907. Cham: Springer, 2016, pp. 87–102. doi: 10.1007/978-3-319-46487-9_6 (cit. on p. 60).

[Guo17a] Guo, Chuan; Pleiss, Geoff; Sun, Yu and Weinberger, Kilian Q.: "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. JMLR.org, 2017, pp. 1321–1330 (cit. on p. 42).

[Guo17b] Guo, Hao; Fan, Xiaochuan and Wang, Song: "Human attribute recognition by refining attention heat map". In: *Pattern Recognition Letters* 94 (2017). PII: S0167865517301605, pp. 38–45. doi: 10.1016/j.patrec.2017.05.012 (cit. on p. 33).

[Guo19] Guo, Hao; Zheng, Kang; Fan, Xiaochuan; Yu, Hongkai and Wang, Song: "Visual attention consistency under image transforms for multi-label image classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 729–739 (cit. on pp. 33, 190, 192, 193).

[Hal18] Halstead, Michael; Denman, Simon; Fookes, Clinton; Tian, YingLi and Nixon, Mark S.: "Semantic Person Retrieval in Surveillance Using Soft Biometrics: AVSS 2018 Challenge II". In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2018). IEEE, 2018. doi: 10.1109/AVSS.2018.8639379 (cit. on p. 54).

[Han19] Han, Kai; Wang, Yunhe; Shu, Han; Liu, Chuanjian; Xu, Chunjing and Xu, Chang: "Attribute aware pooling for pedestrian attribute recognition". In: *IJCAI*. 2019 (cit. on p. 30).

[Han21] Han, Xiaoping; He, Sen; Zhang, Li and Xiang, Tao: "Text-Based Person Search with Limited Data". In: *British Machine Vision Conference*. 2021 (cit. on p. 28).

[Har18] Hara, Kensho; Kataoka, Hirokatsu and Satoh, Yutaka: "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 6546–6555 (cit. on pp. 36, 127, 129, 131, 132).

[He15] He, Kaiming and Sun, Jian: "Convolutional neural networks at constrained time cost". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). Boston, Massachusetts, USA, 7 - 12 June 2015*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, MA, USA). Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2015, pp. 5353–5360. DOI: 10.1109/CVPR.2015.7299173 (cit. on p. 77).

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing and Sun, Jian: "Deep Residual Learning for Image Recognition". In: *29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA). Ed. by Bajcsy, Růžena; Li, Fei-Fei and Tuytelaars, Tinne. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 28, 77, 86, 87).

[He20] He, Y.; Han, J.; Yu, W.; Hong, X.; Wei, X. and Gong, Y.: "City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching". In: *CVPR*. 2020, pp. 576–577 (cit. on p. 200).

[Hir11] Hirzer, Martin; Beleznai, Csaba; Roth, Peter M. and Bischof, Horst: "Person Re-identification by Descriptive and Discriminative Classification". In: *Image analysis. 17th Scandinavian conference, SCIA 2011, Ystad, Sweden, May 2011 ; proceedings*. Ed. by Heyden, Anders and Kahl, Fredrik. Vol. 6688. Lecture Notes in Computer Science / Image Processing, Computer Vision, Pattern Recognition, and Graphics 6688. Berlin and Heidelberg: Springer, 2011, pp. 91–102. DOI: 10.1007/978-3-642-21227-7_9 (cit. on pp. 53, 54).

[Hoc97] Hochreiter, Sepp and Schmidhuber, J.: "Long short-term memory". eng. In: *Neural Computation* 9.8 (1997). Journal Article Research Support, Non-U.S. Gov't, pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735. eprint: 9377276 (cit. on p. 27).

[Hof17] Hoffer, Elad; Hubara, Itay and Soudry, Daniel: "Train longer, generalize better: closing the generalization gap in large batch training of neural networks". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 99, 103, 104).

[How17] Howard, Andrew G.; Zhu, Menglong; Chen, Bo; Kalenichenko, Dmitry; Wang, Weijun; Weyand, Tobias; Andreetto, Marco and Adam, Hartwig: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. 2017. doi: 10.48550/arXiv.1704.04861 (cit. on p. 28).

[HTF56] HTF Rhodes: Alphonse Bertillon: Father of scientific detection. Rhodes HTF (1956) Alphonse Bertillon: Father of scientific detection. Abelard-Schuman, New York. 1956 (cit. on p. 18).

[Iod20] Iodice, Sara and Mikolajczyk, Krystian: "Text Attribute Aggregation and Visual Feature Decomposition for Person Search". In: *BMVC*. 2020 (cit. on pp. 37, 191).

[Iof15] Ioffe, Sergey and Szegedy, Christian: "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456 (cit. on pp. 40, 117).

[Izm18] Izmailov, Pavel; Podoprikhin, Dmitrii; Garipov, T.; Vetrov, Dmitry P. and Wilson, Andrew Gordon: "Averaging Weights Leads to Wider Optima and Better Generalization". In: *Conference on Uncertainty in Artificial Intelligence*. 2018 (cit. on p. 90).

[Jai04a] Jain, Anil K.; Dass, Sarat C. and Nandakumar, Karthik: "Can soft biometric traits assist user recognition?" In: *Biometric Technology for Human Identification*. Defense and Security (Orlando, FL). Ed. by Jain, Anil K. and Ratha, Nalini K. Proceedings of SPIE. SPIE, 2004, pp. 561–572. doi: 10.1117/12.542890 (cit. on pp. 4, 18, 20).

[Jai04b] Jain, Anil K.; Dass, Sarat C. and Nandakumar, Karthik: "Soft Biometric Traits for Personal Recognition Systems". In: *Biometric authentication. First international conference, ICBA 2004, Hong Kong, China, July 15 - 17, 2004 ; proceedings*. Ed. by Zhang, David and Jain, Anil K. Vol. 3072. Lecture notes in computer science 3072. Berlin and Heidelberg: Springer, 2004, pp. 731–738. doi: 10.1007/978-3-540-25948-0_99 (cit. on pp. 20, 21).

[Jai08] Jain, Anil K.; Flynn, Patrick and Ross, Arun A., eds.: Handbook of Biometrics. eng. SpringerLink Bücher. Boston, MA: Springer US, 2008. doi: 10.1007/978-0-387-71041-9 (cit. on pp. 17, 18).

[Jeo21] Jeong, Boseung; Park, Jicheol and Kwak, Suha: "ASMR: Learning Attribute-Based Person Search with Adaptive Semantic Margin Regularizer". In: *2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021 : 11-17 October 2021, virtual event : proceedings*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Montreal, QC, Canada). Ed. by Mortensen, Eric. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 11996–12005. doi: 10.1109/ICCV48922.2021.01180 (cit. on pp. 37, 38, 46, 77, 191).

[Ji20] Ji, Zhong; Hu, Zhenfei; He, Erlu; Han, Jungong and Pang, Yanwei: "Pedestrian attribute recognition based on multiple time steps attention". In: *Pattern Recognition Letters* 138 (2020). PII: S0167865520302646, pp. 170–176. doi: 10.1016/j.patrec.2020.07.018 (cit. on pp. 34, 95).

[Ji22] Ji, Zhong; Hu, Zhenfei; Wang, Yaodong and Li, Shengjia: Reinforced Pedestrian Attribute Recognition with Group Optimization Reward. 2022. url: https://arxiv.org/pdf/2205.14042.pdf (cit. on pp. 29, 77, 190).

[Jia21a] Jia, Jian; Chen, Xiaotang and Huang, Kaiqi: "Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition". In: *2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021 : 11-17 October 2021, virtual event*

: *proceedings*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Montreal, QC, Canada). Ed. by Mortensen, Eric. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 942–951. doi: 10.1109/ICCV48922.2021.00100 (cit. on pp. 33, 190).

[Jia21b] Jia, Jian; Huang, Houjing; Chen, Xiaotang and Huang, Kaiqi: Rethinking of Pedestrian Attribute Recognition: A Reliable Evaluation under Zero-Shot Pedestrian Identity Setting. arXiv, 2021. doi: 10.48550/arXiv.2107.03576 (cit. on pp. 30, 39, 48, 75, 76, 78–80, 83, 84, 95, 107, 190–193).

[Jia23] Jiang, Ding and Ye, Mang: Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. Accepted by CVPR 2023. Codes are available at this https://github.com/anosorae/IRRA. 2023. url: https://arxiv.org/pdf/2303.12501 (cit. on p. 28).

[Joc23] Jocher, Glenn; Chaurasia, Ayush and Qiu, Jing: YOLO by Ultralytics. 1. 2023. url: https://github.com/ultralytics/ultralytics (cit. on p. 200).

[Kat20] Kataoka, Hirokatsu; Wakamiya, Tenga; Hara, Kensho and Satoh, Yutaka: "Would mega-scale datasets further enhance spatiotemporal 3D CNNs?" In: *arXiv preprint arXiv:2004.04968* (2020) (cit. on p. 130).

[Kes16] Keskar, Nitish Shirish; Mudigere, Dheevatsa; Nocedal, Jorge; Smelyanskiy, Mikhail and Tang, Ping Tak Peter: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. Accepted as a conference paper at ICLR 2017. 2016. url: https://arxiv.org/pdf/1609.04836.pdf (cit. on pp. 99, 103).

[Kin14] Kingma, Diederik P. and Ba, Jimmy: "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 106, 113).

[Köh20] Köhl, Philipp; Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "The MTA Dataset for Multi-Target Multi-Camera Pedestrian Tracking by Weighted Distance Aggregation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2020, Virtual, Online, United States, 14 - 19 June 2020*. Institute of Electrical and Electronics Engineers (IEEE), 2020. DOI: 10.1109/CVPRW50498.2020.00529 (cit. on pp. 16, 51, 60, 197, 198, 200, 202).

[Kol20] Kolesnikov, Alexander; Beyer, Lucas; Zhai, Xiaohua; Puigcerver, Joan; Yung, Jessica; Gelly, Sylvain and Houlsby, Neil: "Big Transfer (BiT): General Visual Representation Learning". In: *Computer Vision – ECCV 2020. 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*. Ed. by Vedaldi, Andrea; Bischof, Horst; Brox, Thomas and Frahm, Jan-Michael. 1st ed. 2020. Vol. 12350. Springer eBook Collection 12350. Cham: Springer International Publishing and Imprint Springer, 2020, pp. 491–507. DOI: 10.1007/978-3-030-58558-7_29 (cit. on p. 78).

[Kue74] Kuehn, Lowell L.: "Looking down a Gun Barrel: Person Perception and Violent Crime". In: *Perceptual and Motor Skills* 39.3 (1974), pp. 1159–1164. DOI: 10.2466/pms.1974.39.3.1159 (cit. on p. 22).

[Kum21] Kumar, S. V. Aruna; Yaghoubi, Ehsan; Das, Abhijit; Harish, B. S. and Proenca, Hugo: "The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices". In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 1696–1708. DOI: 10.1109/TIFS.2020.3040881 (cit. on p. 54).

[Kuo10] Kuo, Cheng-Hao; Huang, Chang and Nevatia, Ram: "Inter-camera association of multi-target tracks by on-line learned appearance affinity models". In: *European Conference on Computer Vision*. Springer. 2010, pp. 383–396 (cit. on p. 198).

[Lan22] Lange, Matthias de; Aljundi, Rahaf; Masana, Marc; Parisot, Sarah; Jia, Xu; Leonardis, Ales; Slabaugh, Gregory and Tuytelaars, Tinne: "A Continual Learning Survey: Defying Forgetting in Classification Tasks". eng. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.7 (2022). Journal Article Research Support, Non-U.S. Gov't Journal Article Research Support, Non-U.S. Gov't, pp. 3366–3385. doi: 10.1109/TPAMI.2021.3057446. eprint: 33544669 (cit. on p. 211).

[Lay12] Layne, Ryan; Hospedales, Timothy M. and Gong, Shaogang: "Person Re-identification by Attributes". In: Procedings of the British Machine Vision Conference 2012. British Machine Vision Association, 2012. doi: 10.5244/C.26.24 (cit. on pp. 53–55).

[Lay14] Layne, Ryan; Hospedales, Timothy M. and Gong, Shaogang: Attributes-Based Re-identification. 2014. doi: 10.1007/978-1-4471-6296-4_5 (cit. on pp. 53, 54).

[Li15] Li, Dangwei; Chen, Xiaotang and Huang, Kaiqi: "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios". In: *Third IAPR Asian Conference on Pattern Recognition - ACPR 2015. 3-6 November 2015, Kuala Lumpur, Malaysia : proceedings.* 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) (Kuala Lumpur, Malaysia). International Association for Pattern Recognition. Piscataway, NJ: IEEE, 2015, pp. 111–115. doi: 10.1109/ACPR.2015.7486476 (cit. on pp. 28, 29, 34, 79, 145, 154).

[Li16a] Li, Dangwei; Zhang, Zhang; Chen, Xiaotang; Ling, Haibin and Huang, Kaiqi: A Richly Annotated Dataset for Pedestrian Attribute Recognition. 16 pages, 8 figures. 2016. url: https://arxiv.org/pdf/1603.07054 (cit. on pp. 30, 48, 57, 66).

[Li16b] Li, Yining; Huang, Chen; Loy, Chen Change and Tang, Xiaoou: "Human attribute recognition by deep hierarchical contexts". In:

Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. Springer. 2016, pp. 684–700 (cit. on p. 31).

[Li17a] Li, Shuang; Xiao, Tong; Li, Hongsheng; Yang, Wei and Wang, Xiaogang: "Identity-Aware Textual-Visual Matching with Latent Co-attention". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE International Conference on Computer Vision (ICCV) (2017). IEEE, 2017. DOI: 10.1109/ICCV.2017.209 (cit. on pp. 27, 28).

[Li17b] Li, Shuang; Xiao, Tong; Li, Hongsheng; Zhou, Bolei; Yue, Dayu and Wang, Xiaogang: "Person Search with Natural Language Description". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. DOI: 10.1109/CVPR.2017.551 (cit. on p. 27).

[Li18] Li, Dangwei; Chen, Xiaotang; Zhang, Zhang and Huang, Kaiqi: "Pose Guided Deep Model for Pedestrian Attribute Recognition in Surveillance Scenarios". In: *2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego : 23-27 July 2018*. 2018 IEEE International Conference on Multimedia and Expo (ICME) (San Diego, CA). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2018, pp. 1–6. DOI: 10.1109/ICME.2018.8486604 (cit. on p. 32).

[Li19a] Li, Dangwei; Zhang, Zhang; Chen, Xiaotang and Huang, Kaiqi: "A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios". eng. In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 28.4 (2019). Journal Article, pp. 1575–1590. DOI: 10.1109/TIP.2018.2878349. eprint: 30371372 (cit. on pp. 9, 11, 12, 48, 53, 54, 57, 60, 134, 141–143, 184, 189).

[Li19b] Li, Qiaozhe; Zhao, Xin; He, Ran and Huang, Kaiqi: "Visual-Semantic Graph Reasoning for Pedestrian Attribute Recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (2019), pp. 8634–8641. DOI: 10.1609/aaai.v33i01.33018634 (cit. on p. 29).

[Li22] Li, Shiping; Cao, Min and Zhang, Min: "Learning Semantic-Aligned Feature Representation for Text-Based Person Search". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Singapore, Singapore). IEEE, 2022, pp. 2724–2728. DOI: 10.1109/ICASSP43922.2022.9746846 (cit. on p. 28).

[Lin17] Lin, Tsung-Yi; Goyal, Priya; Girshick, Ross; He, Kaiming and Dollár, Piotr: "Focal Loss for Dense Object Detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE International Conference on Computer Vision (ICCV) (2017). IEEE, 2017, pp. 2980–2988 (cit. on pp. 95, 96, 113, 146).

[Lin19] Lin, Yutian; Zheng, Liang; Zheng, Zhedong; Wu, Yu; Hu, Zhilan; Yan, Chenggang and Yang, Yi: "Improving Person Re-identification by Attribute and Identity Learning". In: *Pattern Recognition* 95 (2019). Accepted to Pattern Recognition (PR), pp. 151–161. DOI: 10.1016/j.patcog.2019.06.006. URL: https://arxiv.org/pdf/1703.07220.pdf (visited on 01/21/2020) (cit. on pp. 11, 25, 46, 53, 54, 56, 179, 180, 182–184, 189).

[Lin94] Lindsay, R. C. L.; Martin, Ronald and Webber, Lisa: "Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy". In: *Law and human behavior* 18.5 (1994), pp. 527–541. DOI: 10.1007/BF01499172 (cit. on pp. 22, 23, 60).

[Liu17] Liu, Xihui; Zhao, Haiyu; Tian, Maoqing; Sheng, Lu; Shao, Jing; Yi, Shuai; Yan, Junjie and Wang, Xiaogang: "HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis". In: *2017 IEEE International Conference on Computer Vision. ICCV 2017 : proceedings : 22 - 29 October 2017, Venice, Italy*. 2017 IEEE International Conference on Computer Vision (ICCV) (Venice). Ed. by Ikeuchi, Katsushi. IEEE Xplore Digital Library. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2017, pp. 350–359. doi: 10.1109/ICCV.2017.46 (cit. on pp. 7, 10–12, 32, 48, 53, 54, 57, 60, 184, 186, 188, 189).

[Liu18a] Liu, Hao; Wu, Jingjing; Jiang, Jianguo; Qi, Meibin and Ren, Bo: Sequence-based Person Attribute Recognition with Joint CTC-Attention Model. 2018. url: https://arxiv.org/pdf/1811.08115 (cit. on p. 29).

[Liu18b] Liu, Pengze; Liu, Xihui; Yan, Junjie and Shao, Jing: "Localization guided learning for pedestrian attribute recognition". In: *arXiv preprint arXiv:1808.09102* (2018) (cit. on p. 32).

[Liu19] Liu, Liyuan; Jiang, Haoming; He, Pengcheng; Chen, Weizhu; Liu, Xiaodong; Gao, Jianfeng and Han, Jiawei: "On the variance of the adaptive learning rate and beyond". In: *arXiv preprint arXiv:1908.03265* (2019) (cit. on p. 107).

[Liu21] 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 2021 : 11-17 October 2021, virtual event : proceedings*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Montreal, QC, Canada). Ed. by Mortensen, Eric. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986 (cit. on pp. 86, 87, 113).

[Liu22] Liu, Zhuang; Mao, Hanzi; Wu, Chao-Yuan; Feichtenhofer, Christoph; Darrell, Trevor and Xie, Saining: "A ConvNet for

the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (New Orleans, LA, USA). IEEE Computer Society, 6/18/2022 - 6/24/2022, pp. 11966–11976. DOI: 10.1109/CVPR52688.2022.01167 (cit. on pp. 86–88, 113).

[Los17] Loshchilov, Ilya and Hutter, Frank: "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2017. URL: https://api.semanticscholar.org/CorpusID:53592270 (cit. on pp. 107, 113).

[Loy10] Loy, Chen Change; Xiang, Tao and Gong, Shaogang: "Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding". In: *International Journal of Computer Vision* 90.1 (2010). PII: 347, pp. 106–129. DOI: 10.1007/s11263-010-0347-5 (cit. on pp. 53, 54).

[Lu23] Lu, Wei-Qing; Hu, Hai-Miao; Yu, Jinzuo; Zhou, Yibo; Wang, Hanzi and Li, Bo: "Orientation-Aware Pedestrian Attribute Recognition based on Graph Convolution Network". In: *IEEE Transactions on Multimedia* (2023), pp. 1–13. DOI: 10.1109/TMM.2023.3259686 (cit. on p. 29).

[Luc18] Lucena, Brian: "Spline-based probability calibration". In: *arXiv preprint arXiv:1809.07751* (2018) (cit. on pp. 50, 157, 161–163).

[Mad18] Madras, David; Pitassi, Toniann and Zemel, Richard: "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc, 2018, pp. 6150–6160 (cit. on p. 42).

[Mad19] Maddox, Wesley J.; Garipov, Timur; Izmailov, Pavel; Vetrov, Dmitry and Wilson, Andrew Gordon: "A Simple Baseline for Bayesian Uncertainty in Deep Learning". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc, 2019 (cit. on p. 42).

[Mah21] Mahdavi, Atefeh and Carvalho, Marco: "A Survey on Open Set Recognition". In: *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering. AIKE 2021 : 1-3 December 2021, Laguna Hills, USA : proceedings*. 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (Laguna Hills, CA, USA). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2021, pp. 37–44. DOI: 10.1109/AIKE52691.2021.00013 (cit. on p. 42).

[Man09] Manning, Christopher D.; Raghavan, Prabhakar and Schütze, Hinrich: Introduction to information retrieval. eng. Reprinted. Cambridge: Cambridge Univ. Press, 2009. 482 pp. (cit. on p. 66).

[Mar16] Martinho-Corbishley, Daniel; Nixon, Mark S. and Carter, John N.: "Soft biometric retrieval to describe and identify surveillance images". In: 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA) (2016). IEEE, 2016. DOI: 10.1109/ISBA.2016.7477240 (cit. on pp. 53, 54).

[McL16] McLaughlin, Niall; Martinez del Rincon, Jesus and Miller, Paul: "Recurrent Convolutional Network for Video-Based Person Re-identification". In: *29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA). Ed. by Bajcsy, Růžena; Li, Fei-Fei and Tuytelaars, Tinne. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2016, pp. 1325–1334. DOI: 10.1109/CVPR.2016.148 (cit. on p. 36).

[Mei07] Meissner, Christain A.; Sporer, Siegfried L. and Schooler, Jonathan W.: "Person Descriptions as Eyewitness Evidence". In: *The Handbook of Eyewitness Psychology: Volume II*. Ed. by Lindsay, R.C.L.; Ross, David F.; Read, J. Don and Toglia, Michael P. Psychology Press, 2007, pp. 1–34 (cit. on pp. 21, 22).

[Mik13] Mikolov, Tomas; Chen, Kai; Corrado, Greg and Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. 2013. url: https://arxiv.org/pdf/1301.3781.pdf (cit. on p. 191).

[Moz20] Mozannar, Hussein and Sontag, David: "Consistent Estimators for Learning to Defer to an Expert". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020 (cit. on p. 42).

[Mül19] Müller, Rafael; Kornblith, Simon and Hinton, Geoffrey E.: "When does label smoothing help?" In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 85).

[Mün15] Münsterberg, Hugo: On the witness stand: Essays in psychology and crime. New York: Doubleday, Page & Company, 1915. doi: 10.1037/10854-000 (cit. on p. 24).

[Nap22] Naphade, Milind et al.: "The 6th AI City Challenge". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (New Orleans, LA, USA). IEEE, 2022, pp. 3346–3355. doi: 10.1109/CVPRW56347.2022.00378 (cit. on p. 2).

[Nei21] Neimark, Daniel; Bar, Omri; Zohar, Maya and Asselmann, Dotan: "Video Transformer Network". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops. ICCVW 2021 : 11-17 October 2021, virtual event : proceedings*. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (Montreal, BC, Canada). Ed. by Berg, Tamara. Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2021, pp. 3156–3165. doi: 10.1109/ICCVW54120.2021.00355 (cit. on pp. 36, 39, 127, 129, 131, 132).

[Niu20] Niu, Kai; Huang, Yan; Ouyang, Wanli and Wang, Liang: "Improving Description-based Person Re-identification by Multi-granularity Image-text Alignments". eng. In: *IEEE Transactions on Image Processing* 29 (2020). Journal Article, pp. 5542–5556. doi: 10.1109/TIP.2020.2984883. eprint: 32275593 (cit. on p. 28).

[Pen14] Pennington, Jeffrey; Socher, Richard and Manning, Christopher: "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar). Ed. by Alessandro Moschitti, Qatar Computing Research Institute; Bo Pang, Google and Walter Daelemans, University of Antwerp. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162 (cit. on p. 191).

[Pla99] Platt, John: "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74 (cit. on pp. 50, 157, 161–163).

[Qia19] Qiao, Siyuan; Wang, Huiyu; Liu, Chenxi; Shen, Wei and Yuille, Alan: Micro-Batch Training with Batch-Channel Normalization and Weight Standardization. 2019. url: https://arxiv.org/pdf/1903.10520.pdf (cit. on p. 40).

[Rad21] Radford, Alec et al.: Learning Transferable Visual Models From Natural Language Supervision. 2021. doi: 10.48550/arXiv.2103.00020 (cit. on p. 28).

[Rah22] Rahman, Quazi Marufur; Sunderhauf, Niko; Corke, Peter and Dayoub, Feras: "FSNet: A Failure Detection Framework for Semantic Segmentation". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 3030–3037. doi: 10.1109/LRA.2022.3143219 (cit. on pp. 42, 135).

[Ram18] Ramanagopal, Manikandasriram Srinivasan; Anderson, Cyrus; Vasudevan, Ram and Johnson-Roberson, Matthew: "Failing to Learn: Autonomously Identifying Perception Failures for Self-Driving Cars". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 3860–3867. doi: 10.1109/LRA.2018.2857402 (cit. on pp. 42, 135).

[Rei11] Reid, Daniel A. and Nixon, Mark S.: "Using comparative human descriptions for soft biometrics". In: *International Joint Conference on Biometrics (IJCB), 2011. 11 - 13 Oct. 2011, Washington, DC, USA ; [a special combination of two biometrics research conferences: IAPR International Conference on Biometrics (ICB) and IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2011 IEEE International Joint Conference on Biometrics (IJCB) (Washington, DC, USA). International Association for Pattern Recognition and Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2011, pp. 1–6. DOI: 10.1109/IJCB.2011.6117513 (cit. on pp. 4, 18, 21).

[Ren16] Ren, Shaoqing; He, Kaiming; Girshick, Ross and Sun, Jian: "Faster R-CNN: towards real-time object detection with region proposal networks". In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149 (cit. on p. 200).

[Ric23] Richter, Jan-Philip; Flores, Sebastian and Urbann, Oliver: "Online Object Tracking on Multiple Cameras with Completely Overlapping Views". In: *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*. 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE) (Helsinki, Finland). IEEE, 2023, pp. 1–7. DOI: 10.1109/ISIE51358.2023.10228087 (cit. on p. 203).

[Ris16] Ristani, Ergys; Solera, Francesco; Zou, Roger S.; Cucchiara, Rita and Tomasi, Carlo: "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking". In: *Computer vision - ECCV 2016 Workshops. Amsterdam, The Netherlands, October 8-10 and 15-16, 2016 : proceedings*. Ed. by Hua, Gang and Jégou, Hervé. Vol. 9914. Lecture notes in computer science 9914. Cham: Springer, 2016, pp. 17–35. DOI: 10.1007/978-3-319-48881-3_2 (cit. on pp. 60, 198, 201).

[Ris18] Ristani, Ergys and Tomasi, Carlo: "Features for multi-target multi-camera tracking and re-identification". In: *Proceedings of*

the IEEE conference on computer vision and pattern recognition. 2018, pp. 6036–6046 (cit. on p. 200).

[Rus15] Russakovsky, Olga et al.: "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (2015). PII: 816, pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cit. on pp. 78, 80).

[Sam08] Samangooei, Sina; Guo, Baofeng and Nixon, Mark S.: "The Use of Semantic Human Description as a Soft Biometric". In: 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems. IEEE, 2008. DOI: 10.1109/BTAS.2008.4699354 (cit. on pp. 4, 18, 21).

[Sar17] Sarfraz: "Deep view-sensitive pedestrian attribute inference in an end-to-end model". In: *arXiv:1707.06089* (2017) (cit. on pp. 33, 38).

[Sar18a] Sarafianos, Nikolaos; Giannakopoulos, Theodoros; Nikou, Christophoros and Kakadiaris, Ioannis A.: "Curriculum learning of visual attribute clusters for multi-task classification". In: *Pattern Recognition* 80 (2018). PII: S0031320318300840, pp. 94–108. DOI: 10.1016/j.patcog.2018.02.028 (cit. on p. 29).

[Sar18b] Sarafianos, Nikolaos; Xu, Xiang and Kakadiaris, Ioannis A.: "Deep Imbalanced Attribute Classification Using Visual Attention Aggregation". In: *Computer vision - ECCV 2018. 15th European conference, Munich, Germany, September 8-14, 2018 : proceedings*. Ed. by Ferrari, Vittorio; Hebert, Martial; Sminchisescu, Cristian and Weiss, Yair. Lecture notes in computer science 11215. Cham: Springer, 2018, pp. 708–725. DOI: 10.1007/978-3-030-01252-6_42 (cit. on pp. 33, 38, 95, 190).

[Sar19] Sarafianos, Nikolaos; Xu, Xiang and Kakadiaris, Ioannis A.: "Adversarial Representation Learning for Text-to-Image Matching". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019). IEEE, 2019. DOI: 10.1109/ICCV.2019.00591 (cit. on p. 28).

[Sat20] Satariano, Adam: "London Police Are Taking Surveillance to a Whole New Level". In: *The New York Times* (Jan. 24, 2020). URL: https://www.nytimes.com/2020/01/24/business/london-police-facial-recognition.html (visited on 11/24/2022) (cit. on pp. 2, 20, 21).

[Sch18] Schumann, Arne; Specker, Andreas and Beyerer, Jürgen: "Attribute-based Person Retrieval and Search in Video Sequences". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (Auckland, New Zealand). IEEE, 2018, pp. 1–6. DOI: 10.1109/AVSS.2018.8639114 (cit. on pp. 14, 79, 84).

[Sha22] Shao, Zhiyin; Zhang, Xinyu; Fang, Meng; Lin, Zhifeng; Wang, Jian and Ding, Changxing: "Learning Granularity-Unified Representations for Text-to-Image Person Re-identification". In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22: The 30th ACM International Conference on Multimedia (Lisboa Portugal). Ed. by Magalhães, João; Del Bimbo, Alberto; Satoh, Shin'ichi; Sebe, Nicu; Alameda-Pineda, Xavier; Jin, Qin; Oria, Vincent and Toni, Laura. New York, NY, USA: ACM, 2022, pp. 5566–5574. DOI: 10.1145/3503161.3548028 (cit. on p. 28).

[Shu23] Shu, Xiujun; Wen, Wei; Wu, Haoqian; Chen, Keyu; Song, Yiran; Qiao, Ruizhi; Ren, Bo and Wang, Xiao: "See Finer, See More: Implicit Modality Alignment for Text-Based Person Retrieval". In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Karlinsky, Leonid; Michaeli, Tomer and Nishino, Ko. Vol. 13805. Lecture notes in computer science. Cham: Springer Nature Switzerland, 2023, pp. 624–641. DOI: 10.1007/978-3-031-25072-9_42 (cit. on p. 28).

[Sim14] Simonyan, Karen and Zisserman, Andrew: "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 27, 77).

[Smi17] Smith, Samuel L.; Kindermans, Pieter-Jan; Ying, Chris and Le V, Quoc: Don't Decay the Learning Rate, Increase the Batch Size. 11 pages, 8 figures. Published as a conference paper at ICLR 2018. 2017. URL: https://arxiv.org/pdf/1711.00489.pdf (cit. on p. 99).

[Spe20a] Specker, Andreas: "A Realistic Predictor for Pedestrian Attribute Recognition". In: *Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Hrsg.: J. Beyerer; T. Zander.* Vol. 45. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2020, pp. 95–112 (cit. on pp. 14, 49, 135).

[Spe20b] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "A multitask model for person re-identification and attribute recognition using semantic regions". In: *Artificial Intelligence and Machine Learning in Defense Applications II. Ed.: J. Dijk.* Vol. 11543. Proceedings of SPIE. SPIE, 2020, p. 115430I. DOI: 10.1117/12.2573981 (cit. on pp. 14, 25, 46, 78, 84).

[Spe20c] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "An Evaluation Of Design Choices For Pedestrian Attribute Recognition In Video". In: *2020 IEEE International Conference on Image Processing. Proceedings : September 25-28, 2020 : virtual conference, Abu Dhabi, United Arab Emirates.* 2020 IEEE International Conference on Image Processing (ICIP) (Abu Dhabi, United Arab Emirates). Institute of Electrical and Electronics Engineers and IEEE Signal Processing Society. Piscataway, NJ: IEEE, 2020, pp. 2331–2335. DOI: 10.1109/ICIP40778.2020.9191264 (cit. on pp. 14, 49, 84, 125).

[Spe21a] Specker, Andreas and Beyerer, Jürgen: "Improving Attribute-Based Person Retrieval By Using A Calibrated, Weighted, And Distribution-Based Distance Metric". In: *2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, Anchorage, AK, USA, 19-22 Sept. 2021*. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 2378–2382. DOI: 10.1109/ICIP42928.2021.9506096 (cit. on pp. 15, 51, 155).

[Spe21b] Specker, Andreas; Stadler, Daniel; Florin, Lucas and Beyerer, Jürgen: "An Occlusion-Aware Multi-Target Multi-Camera Tracking System". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 4168–4177. DOI: 10.1109/CVPRW53098.2021.00471 (cit. on p. 198).

[Spe22a] Specker, Andreas and Beyerer, Jürgen: "Toward Accurate Online Multi-target Multi-camera Tracking in Real-time". In: *30th European Signal Processing Conference (EUSIPCO 2022). Proceedings : 29 August-2 September 2022, Belgrade, Serbia*. 2022 30th European Signal Processing Conference (EUSIPCO) (Belgrade, Serbia). Ed. by Routtenberg, Tirza and Tadić, Predrag. European Association for Signal Processing. Piscataway, NJ: IEEE, 2022, pp. 533–537. DOI: 10.23919/EUSIPCO55093.2022.9909670 (cit. on pp. 16, 51, 197, 198, 202).

[Spe22b] Specker, Andreas; Florin, Lucas; Cormier, Mickael and Beyerer, Jürgen: "Improving Multi-Target Multi-Camera Tracking by Track Refinement and Completion". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (New Orleans, LA, USA). IEEE, 2022, pp. 3198–3208. DOI: 10.1109/CVPRW56347.2022.00361 (cit. on p. 198).

[Spe22c] Specker, Andreas; Moritz, Lennart; Cormier, Mickael and Beyerer, Jürgen: "Fast and Lightweight Online Person Search

for Large-Scale Surveillance Systems". In: *2022 IEEE Winter Conference on Applications of Computer Vision Workshops. WACVW 2022 : 4-8 January 2022, Waikoloa, Hawaii*. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2022, pp. 570–580. DOI: 10.1109/WACVW54805.2022.00063 (cit. on pp. 16, 51, 197, 198, 204, 205).

[Spe23a] Specker, Andreas and Beyerer, Jürgen: "Balanced Pedestrian Attribute Recognition for Improved Attribute-based Person Retrieval". In: *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*. 2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS) (Guayaquil, Ecuador). IEEE, 7/4/2023 - 7/7/2023, pp. 1–7. DOI: 10.1109/ICPRS58416.2023.10178990 (cit. on pp. 14, 15, 48, 49, 64, 116, 154, 211).

[Spe23b] Specker, Andreas; Cormier, Mickael and Beyerer, Jürgen: "UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval". In: *2023 IEEE Winter Conference on Applications of Computer Vision. 3-7 January 2023, Waikoloa, Hawaii : proceedings*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (Waikoloa, HI, USA). Ed. by Berg, Tamara. IEEE Computer Society and Computer Vision Foundation. Piscataway, NJ: IEEE, 2023, pp. 981–990. DOI: 10.1109/WACV56688.2023.00104 (cit. on pp. 14, 15, 48, 49, 59, 79, 84).

[Spi18] Spiegel, Der: "G20-Krawalle in Hamburg: Polizei startet konzertierte Razzien in vier Ländern". In: *DER SPIEGEL* (May 29, 2018). URL: https://www.spiegel.de/panorama/justiz/g20-krawalle-in-hamburg-durchsuchungen-in-vier-laendern-a-1210023.html (visited on 11/01/2023) (cit. on p. 2).

[Spo92] Sporer, Siegfried L.: "An archival analysis of person descriptions". In: *REPORT ON THE BIENNIAL MEETING OF THE AMERICAN PSYCHOLOGY-LAW SOCIETY IN SAN-DIEGO, CALIFORNIA, IN MARCH 1992*. 1992 (cit. on pp. 18, 22–24, 60, 154).

[Sri14] Srivastava, Nitish; Hinton, Geoffrey; Krizhevsky, Alex; Sutskever, Ilya and Salakhutdinov, Ruslan: "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958 (cit. on pp. 104, 113).

[Sri15] Srivastava, Rupesh Kumar; Greff, Klaus and Schmidhuber, Jürgen: "Training Very Deep Networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 2377–2385 (cit. on p. 77).

[Ste92] Steblay, Nancy Mehrkens: "A meta-analytic review of the weapon focus effect". In: *Law and human behavior* 16.4 (1992), pp. 413–424. doi: 10.1007/BF02352267 (cit. on p. 22).

[Sto19] Stone, Adam: "Surveillance's Role in Controlling Crowds". In: *Security Magazine* (Mar. 12, 2019). url: https://www.securitymagazine.com/articles/89968-surveillances-role-in-controlling-crowds (visited on 11/01/2023) (cit. on p. 1).

[Sud15] Sudowe, Patrick; Spitzer, Hannah and Leibe, Bastian: "Person Attribute Recognition with a Jointly-Trained Holistic CNN Model". In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015). IEEE, 2015. doi: 10.1109/ICCVW.2015.51 (cit. on p. 29).

[Sun19] Sun, Rémy and Lampert, Christoph H.: "KS(conf): A Light-Weight Test if a ConvNet Operates Outside of Its Specifications". In: *Pattern Recognition. 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings*. Ed. by Brox, Thomas; Bruhn, Andrés and Fritz, Mario. Vol. 11269.

SpringerLink Bücher 11269. Cham: Springer International Publishing, 2019, pp. 244–259. DOI: 10.1007/978-3-030-12939-2_18 (cit. on p. 42).

[Sze16] Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon and Wojna, Zbigniew: "Rethinking the Inception Architecture for Computer Vision". In: *29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada.* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA). Ed. by Bajcsy, Růžena; Li, Fei-Fei and Tuytelaars, Tinne. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308 (cit. on pp. 108, 113).

[Tan19a] Tan, Mingxing and Le, Quoc V.: "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Chaudhuri, Kamalika and Salakhutdinov, Ruslan. Vol. 97. Proceedings of Machine Learning Research. 09–15 Jun. PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html (cit. on p. 91).

[Tan19b] Tan, Zichang; Yang, Yang; Wan, Jun; Hang, Hanyuan; Guo, Guodong and Li, Stan Z.: "Attention-Based Pedestrian Attribute Analysis". eng. In: *IEEE Transactions on Image Processing* 28.12 (2019). Journal Article, pp. 6126–6140. DOI: 10.1109/TIP.2019.2919199. eprint: 31283504 (cit. on pp. 34, 38, 190).

[Tan19c] Tang, Chufeng; Sheng, Lu; Zhang, Zhao-Xiang and Hu, Xiaolin: "Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019). IEEE, 2019. DOI: 10.1109/iccv.2019.00510 (cit. on pp. 33, 190, 192, 193).

[Tan20] Tan, Zichang; Yang, Yang; Wan, Jun; Guo, Guodong and Li, Stan Z.: "Relation-Aware Pedestrian Attribute Recognition with Graph Convolutional Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (2020), pp. 12055–12062. DOI: 10.1609/aaai.v34i07.6883 (cit. on pp. 29, 79, 190).

[Tan22] Tang, Zengming and Huang, Jun: "DRFormer: Learning dual relations using Transformer for pedestrian attribute recognition". In: *Neurocomputing* 497 (2022). PII: S0925231222005598, pp. 159–169. DOI: 10.1016/j.neucom.2022.05.028 (cit. on pp. 35, 190).

[Tra15] Tran, Du; Bourdev, Lubomir; Fergus, Rob; Torresani, Lorenzo and Paluri, Manohar: "Learning Spatiotemporal Features with 3D Convolutional Networks". In: *2015 IEEE International Conference on Computer Vision. 11-18 December 2015, Santiago, Chile : proceedings*. 2015 IEEE International Conference on Computer Vision (ICCV) (Santiago, Chile). Ed. by Bajcsy, Ruzena; Hager, Greg and Ma, Yi. IEEE International Conference on Computer Vision and Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2015, pp. 4489–4497. DOI: 10.1109/ICCV.2015.510 (cit. on pp. 36, 39).

[Tra18] Tran, Du; Wang, Heng; Torresani, Lorenzo; Ray, Jamie; LeCun, Yann and Paluri, Manohar: "A Closer Look at Spatiotemporal Convolutions for Action Recognition". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2018 : proceedings : 18-22 June 2018, Salt Lake City, Utah*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, UT). Ed. by Brown, Michael S.; Morse, Bryan and Peleg, Shmuel. Institute of Electrical and Electronics Engineers and Computer Vision Foundation. Piscataway, NJ: IEEE, 2018, pp. 6450–6459. DOI: 10.1109/CVPR.2018.00675 (cit. on pp. 36, 39, 125).

[Uly16] Ulyanov, Dmitry; Vedaldi, Andrea and Lempitsky, Victor: Instance Normalization: The Missing Ingredient for Fast Stylization. 27.07.2016. url: https://arxiv.org/pdf/1607.08022 (cit. on pp. 40, 41).

[van97] van Koppen, Peter J. and Lochun, Shara K.: "Portraying perpetrators: The validity of offender descriptions by witnesses". In: *Law and human behavior* 21.6 (1997), pp. 661–685. doi: 10.1023/A:1024812831576 (cit. on pp. 23, 24).

[Vas17] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz and Polosukhin, Illia: "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 125).

[Wan16] Wang, Jiang; Yang, Yi; Mao, Junhua; Huang, Zhiheng; Huang, Chang and Xu, Wei: "CNN-RNN: A Unified Framework for Multi-label Image Classification". In: *29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada.* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA). Ed. by Bajcsy, Růžena; Li, Fei-Fei and Tuytelaars, Tinne. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2016, pp. 2285–2294. doi: 10.1109/CVPR.2016.251 (cit. on p. 125).

[Wan17] Wang, Xin; Luo, Yujia; Crankshaw, Daniel; Tumanov, Alexey; Yu, Fisher and Gonzalez, Joseph E.: IDK Cascades: Fast Deep Learning by Learning not to Overthink. UAI 2018 camera ready. 3.06.2017. url: https://arxiv.org/pdf/1706.00885.pdf (cit. on p. 42).

[Wan18a] Wang, Pei and Vasconcelos, Nuno: "Towards Realistic Predictors". In: *Computer vision - ECCV 2018. 15th European conference, Munich, Germany, September 8-14, 2018 : proceedings.* Ed. by Ferrari, Vittorio; Hebert, Martial; Sminchisescu, Cristian and Weiss, Yair. Lecture notes in computer science 11217. Cham:

Springer, 2018, pp. 37–53. DOI: 10.1007/978-3-030-01261-8_3 (cit. on pp. 42, 135, 136, 138, 145).

[Wan18b] Wang, Xiaolong; Girshick, Ross; Gupta, Abhinav and He, Kaiming: "Non-local Neural Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. DOI: 10.1109/cvpr.2018.00813 (cit. on pp. 30, 36).

[Wan19a] Wang, Yiru; Gan, Weihao; Yang, Jie; Wu, Wei and Yan, Junjie: "Dynamic Curriculum Learning for Imbalanced Data Classification". In: *2019 International Conference on Computer Vision. ICCV 2019 : proceedings : 27 October-2 November 2019, Seoul, Korea*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (Seoul, Korea (South)). Institute of Electrical and Electronics Engineers and Computer Vision Foundation. Piscataway, NJ: IEEE, 2019, pp. 5016–5025. DOI: 10.1109/ICCV.2019.00512 (cit. on p. 29).

[Wan19b] Wang, Yuyu; Bo, Chunjuan; Wang, Dong; Wang, Shuang; Qi, Yunwei and Lu, Huchuan: "Language Person Search with Mutually Connected Classification Loss". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019). IEEE, 2019. DOI: 10.1109/ICASSP.2019.8682456 (cit. on p. 28).

[Wan20] Wang, Zhe; Fang, Zhiyuan; Wang, Jun and Yang, Yezhou: "ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language". In: *Computer Vision – ECCV 2020. 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*. Ed. by Vedaldi, Andrea; Bischof, Horst; Brox, Thomas and Frahm, Jan-Michael. 1st ed. 2020. Vol. 12357. Springer eBook Collection 12357. Cham: Springer International Publishing and Imprint Springer, 2020, pp. 402–420. DOI: 10.1007/978-3-030-58610-2_24 (cit. on p. 28).

[Wan21] Wang, Wenhai; Xie, Enze; Li, Xiang; Fan, Deng-Ping; Song, Kaitao; Liang, Ding; Lu, Tong; Luo, Ping and Shao, Ling: "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions". In: *2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021 : 11-17 October 2021, virtual event : proceedings*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Montreal, QC, Canada). Ed. by Mortensen, Eric. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 548–558. doi: 10.1109/ICCV48922.2021.00061 (cit. on p. 86).

[Wan22a] Wang, Chien-Yao; Bochkovskiy, Alexey and Liao, Hong-Yuan Mark: "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *arXiv preprint arXiv:2207.02696* (2022) (cit. on p. 200).

[Wan22b] Wang, Wenhai; Xie, Enze; Li, Xiang; Fan, Deng-Ping; Song, Kaitao; Liang, Ding; Lu, Tong; Luo, Ping and Shao, Ling: "PVT v2: Improved baselines with Pyramid Vision Transformer". In: *Computational Visual Media* 8.3 (2022). Computational Visual Media, 2022, Vol. 8, No. 3, Pages: 415-424 Accepted to CVMJ 2022 PII: 274, pp. 415–424. doi: 10.1007/s41095-022-0274-8. url: http://arxiv.org/pdf/2106.13797v6 (cit. on pp. 86, 87).

[Wan22c] Wang, Xiao; Zheng, Shaofei; Yang, Rui; Zheng, Aihua; Chen, Zhe; Tang, Jin and Luo, Bin: "Pedestrian attribute recognition: A survey". In: *Pattern Recognition* 121 (2022). PII: S0031320321004015, p. 108220. doi: 10.1016/j.patcog.2021.108220 (cit. on p. 29).

[Wan22d] Wang, Xinyi; Peng, Jianteng; Zhang, Sufang; Chen, Bihui; Wang, Yi and Guo, Yandong: A Survey of Face Recognition. 2022. url: https://arxiv.org/pdf/2212.13038.pdf (cit. on pp. 4, 21).

[Wan22e] Wang, Zijie; Zhu, Aichun; Xue, Jingyi; Wan, Xili; Liu, Chao; Wang, Tian and Li, Yifeng: "CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval".

In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22: The 30th ACM International Conference on Multimedia (Lisboa Portugal). Ed. by Magalhães, João; Del Bimbo, Alberto; Satoh, Shin'ichi; Sebe, Nicu; Alameda-Pineda, Xavier; Jin, Qin; Oria, Vincent and Toni, Laura. New York, NY, USA: ACM, 2022, pp. 5314–5322. doi: 10.1145/3503161.3548057 (cit. on p. 28).

[Wan22f] Wang, Zijie; Zhu, Aichun; Xue, Jingyi; Wan, Xili; Liu, Chao; Wang, Tian and Li, Yifeng: "Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold". In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22: The 30th ACM International Conference on Multimedia (Lisboa Portugal). Ed. by Magalhães, João; Del Bimbo, Alberto; Satoh, Shin'ichi; Sebe, Nicu; Alameda-Pineda, Xavier; Jin, Qin; Oria, Vincent and Toni, Laura. New York, NY, USA: ACM, 2022, pp. 1984–1992. doi: 10.1145/3503161.3548166 (cit. on p. 28).

[Wan23a] Wang, Cheng: Calibration in Deep Learning: A Survey of the State-of-the-Art. 2.08.2023. url: https://arxiv.org/pdf/2308.01222.pdf (cit. on p. 42).

[Wan23b] Wang, Liyuan; Zhang, Xingxing; Su, Hang and Zhu, Jun: "A comprehensive survey of continual learning: Theory, method and application". In: *arXiv preprint arXiv:2302.00487* (2023) (cit. on p. 211).

[WEL17] WELT: "G20-Führungsstab behält Einsatzgeschehen per Video im Blick". In: *WELT* (June 30, 2017). url: https://www.welt.de/regionales/hamburg/article166117797/G20-Fuehrungsstab-behaelt-Einsatzgeschehen-per-Video-im-Blick.html (visited on 11/01/2023) (cit. on p. 2).

[Wu18] Wu, Yuxin and He, Kaiming: "Group Normalization". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. September. 2018 (cit. on pp. 40, 41).

[Wu21] Wu, Yushuang; Yan, Zizheng; Han, Xiaoguang; Li, Guanbin; Zou, Changqing and Cui, Shuguang: "LapsCore: Language-guided Person Search via Color Reasoning". In: *2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021 : 11-17 October 2021, virtual event : proceedings*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Montreal, QC, Canada). Ed. by Mortensen, Eric. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 1604–1613. DOI: 10.1109/ICCV48922.2021.00165 (cit. on p. 28).

[Xie17] Xie, Saining; Girshick, Ross; Dollar, Piotr; Tu, Zhuowen and He, Kaiming: "Aggregated Residual Transformations for Deep Neural Networks". In: *30th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017 : 21-26 July 2016, Honolulu, Hawaii : proceedings*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI). Ed. by Chellappa, Rama; Zhang, Zhengyou and Hoogs, Anthony. Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634 (cit. on p. 77).

[Yan16] Yang, Luwei; Zhu, Ligeng; Wei, Yichen; Liang, Shuang and Tan, Ping: "Attribute Recognition from Adaptive Parts". In: *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Conference 2016 (York, UK). Ed. by Wilson, R. C.; Hancock, E. R.; Smith, W. A. P.; Pears, N. E. and Bors, A. G. British Machine Vision Association, 2016, pp. 81.1–81.11. DOI: 10.5244/C.30.81 (cit. on p. 32).

[Yan20] Yang, Jie; Fan, Jiarou; Wang, Yiru; Wang, Yige; Gan, Weihao; Liu, Lin and Wu, Wei: "Hierarchical feature embedding for attribute recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13055–13064 (cit. on p. 34).

[Yan22] Yan, Shuanglin; Dong, Neng; Zhang, Liyan and Tang, Jinhui: CLIP-Driven Fine-grained Text-Image Person Re-identification. 2022. URL: https://arxiv.org/pdf/2210.10276 (cit. on p. 28).

[Ye21] Ye, Jin; Yang, Xipeng; Kang, Shuai; He, Yue; Zhang, Weiming; Huang, Leping; Jiang, Minyue; Zhang, Wei; Shi, Yifeng; Xia, Meng et al.: "A robust MTMC tracking system for AI-City Challenge 2021". In: *Proceedings of the IEEE/CVF Conference CVPRW 2021*. 2021, pp. 4044–4053 (cit. on p. 200).

[Ye22] Ye, Mang; Shen, Jianbing; Lin, Gaojie; Xiang, Tao; Shao, Ling and Hoi, Steven C. H.: "Deep Learning for Person Re-Identification: A Survey and Outlook". eng. In: *IEEE transactions on pattern analysis and machine intelligence* 44.6 (2022). Journal Article Review Research Support, Non-U.S. Gov't Journal Article Review Research Support, Non-U.S. Gov't, pp. 2872–2893. DOI: 10.1109/TPAMI.2021.3054775. eprint: 33497329 (cit. on pp. 4, 21, 77).

[Yin18] Yin, Zhou; Zheng, Wei-Shi; Wu, Ancong; Yu, Hong-Xing; Wan, Hai; Guo, Xiaowei; Huang, Feiyue and Lai, Jianhuang: "Adversarial Attribute-Image Person Re-identification". In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (2018). International Joint Conferences on Artificial Intelligence Organization, 2018. DOI: 10.24963/ijcai.2018/153 (cit. on pp. 37, 38, 191).

[Yos18] Yoshikawa, Yuya; Lin, Jiaqing and Takeuchi, Akikazu: "Stair actions: A video dataset of everyday home actions". In: *arXiv preprint arXiv:1804.04326* (2018) (cit. on p. 128).

[Yui86] Yuille, John C. and Cutshall, Judith L.: "A case study of eyewitness memory of a crime". In: *Journal of Applied Psychology* 71.2 (1986), pp. 291–301. DOI: 10.1037/0021-9010.71.2.291 (cit. on pp. 23, 24).

[Zad01] Zadrozny, Bianca and Elkan, Charles: "Obtaining calibrated probability estimates from decision trees and naive bayesian

classifiers". In: *Icml*. Vol. 1. 2001, pp. 609–616 (cit. on pp. 50, 157, 161–163).

[Zha14a] Zhang, Ning; Paluri, Manohar; Ranzato, Marc'Aurelio; Darrell, Trevor and Bourdev, Lubomir: "PANDA: Pose Aligned Networks for Deep Attribute Modeling". In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014). IEEE, 2014. DOI: 10.1109/CVPR.2014.212 (cit. on p. 31).

[Zha14b] Zhang, Peng; Wang, Jiuling; Farhadi, Ali; Hebert, Martial and Parikh, Devi: "Predicting Failures of Vision Systems". In: *CVPR 2014. 2014 IEEE Conference on Computer Vision and Pattern Recognition : proceedings : 23-28 June 2014, Columbus, Ohio*. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Columbus, OH, USA). IEEE Computer Society. Los Alamitos, California: IEEE Computer Society, 2014, pp. 3566–3573. DOI: 10.1109/CVPR.2014.456 (cit. on p. 41).

[Zha15] Zhang, Shu; Staudt, Elliot; Faltemier, Tim and Roy-Chowdhury, Amit K.: "A camera network tracking (camnet) dataset and performance baseline". In: *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE %Master's thesis literature % This file was created with JabRef 2.10. % Encoding: UTF8. 2015, pp. 365–372 (cit. on p. 198).

[Zha18a] Zhang, Ying and Lu, Huchuan: "Deep Cross-Modal Projection Learning for Image-Text Matching". In: *Computer Vision – ECCV 2018*. Ed. by Ferrari, Vittorio; Hebert, Martial; Sminchisescu, Cristian and Weiss, Yair. Vol. 11205. Lecture notes in computer science. Cham: Springer International Publishing, 2018, pp. 707–723. DOI: 10.1007/978-3-030-01246-5_42 (cit. on p. 28).

[Zha18b] Zhao, Xin; Sang, Liufang; Ding, Guiguang; Guo, Yuchen and Jin, Xiaoming: "Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning". In: *IJCAI*. 2018, pp. 3177–3183 (cit. on pp. 41, 117).

[Zha19a] Zhang, Xindi and Izquierdo, Ebroul: "Real-Time Multi-Target Multi-Camera Tracking with Spatial-Temporal Information". In: *2019 IEEE Visual Communications and Image Processing (VCIP)*. 2019, pp. 1–4. doi: 10.1109/VCIP47243.2019. (cit. on p. 203).

[Zha19b] Zhao, Xin; Sang, Liufang; Ding, Guiguang; Han, Jungong; Di, Na and Yan, Chenggang: "Recurrent attention model for pedestrian attribute recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 9275–9282 (cit. on p. 29).

[Zha21a] Zhang, Haoyang; Wang, Ying; Dayoub, Feras and Sunderhauf, Niko: "Varifocalnet: An iou-aware dense object detector". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. IEEE, 2021, pp. 8514–8523 (cit. on p. 200).

[Zha21b] Zhang, Songyang; Li, Zeming; Yan, Shipeng; He, Xuming and Sun, Jian: "Distribution Alignment: A Unified Framework for Long-tail Visual Recognition". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 19-25 June 2021 : proceedings*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Nashville, TN, USA). Ed. by Brown, Michael. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation. Piscataway, NJ: IEEE, 2021, pp. 2361–2370. doi: 10.1109/CVPR46437.2021.00239 (cit. on p. 79).

[Zha22] Zhang, Hang et al.: "ResNeSt: Split-Attention Networks". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (New Orleans, LA, USA). IEEE, 2022, pp. 2735–2745. doi: 10.1109/CVPRW56347.2022.00309 (cit. on p. 77).

[Zhe15] Zheng, Liang; Shen, Liyue; Tian, Lu; Wang, Shengjin; Wang, Jingdong and Tian, Qi: "Scalable Person Re-identification: A

Benchmark". In: *2015 IEEE International Conference on Computer Vision. 11-18 December 2015, Santiago, Chile : proceedings*. 2015 IEEE International Conference on Computer Vision (ICCV) (Santiago, Chile). Ed. by Bajcsy, Ruzena; Hager, Greg and Ma, Yi. IEEE International Conference on Computer Vision and Institute of Electrical and Electronics Engineers (IEEE). Piscataway, NJ: IEEE, 2015, pp. 1116–1124. doi: 10.1109/ICCV.2015.133 (cit. on pp. 5, 7, 8, 10, 12, 48, 50, 53, 54, 56, 60, 179, 180, 182–184, 189).

[Zhe16] Zheng, Liang; Bie, Zhi; Sun, Yifan; Wang, Jingdong; Su, Chi; Wang, Shengjin and Tian, Qi: "Mars: A video benchmark for large-scale person re-identification". In: *European Conference on Computer Vision*. Springer International Publishing. 2016, pp. 868–884 (cit. on pp. 35, 54, 64, 124, 127, 169, 210).

[Zhe20] Zheng, Zhedong; Zheng, Liang; Garrett, Michael; Yang, Yi; Xu, Mingliang and Shen, Yi-Dong: "Dual-path Convolutional Image-Text Embeddings with Instance Loss". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 16.2 (2020), pp. 1–23. doi: 10.1145/3383184 (cit. on p. 28).

[Zhe21] Zheng, Xiaoqiang; Yu, Zhenxia; Chen, Lin; Zhu, Fan and Wang, Shilong: "Multi-label Contrastive Focal Loss for Pedestrian Attribute Recognition". In: *Proceedings of ICPR 2020. 25th International Conference on Pattern Recognition : Milan, 10-15 January 2021*. 2020 25th International Conference on Pattern Recognition (ICPR) (Milan, Italy). Ed. by Vezzani, Roberto. International Association for Pattern Recognition. Piscataway, NJ: IEEE, 2021, pp. 7349–7356. doi: 10.1109/ICPR48806.2021.9411959 (cit. on pp. 34, 95, 190).

[Zho17] Zhou, Yang; Yu, Kai; Leng, Biao; Zhang, Zhang; Li, Dangwei and Huang, Kaiqi: "Weakly-supervised Learning of Mid-level Features for Pedestrian Attribute Recognition and Localization". In: Procedings of the British Machine Vision Conference 2017.

British Machine Vision Association, 2017. DOI: 10.5244/C.31.69 (cit. on p. 34).

[Zho20] Zhou, X.; Koltun, V. and Krähenbühl, P.: "Tracking Objects as Points". In: *ECCV*. 2020, pp. 474–490 (cit. on p. 200).

[Zho21] Zhong, Jiabao; Qiao, Hezhe; Chen, Lin; Shang, Mingsheng and Liu, Qun: "Improving Pedestrian Attribute Recognition with Multi-Scale Spatial Calibration". In: *IJCNN 2021, virtual event, 18-22 July 2021 - the International Joint Conference on Neural Networks. 2021 conference proceedings*. 2021 International Joint Conference on Neural Networks (IJCNN) (Shenzhen, China). International Neural Network Society and IEEE Computational Intelligence Society. Piscataway, NJ, USA: IEEE, 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533647 (cit. on pp. 30, 190).

[Zho23] Zhou, Yibo; Hu, Hai-Miao; Yu, Jinzuo; Xu, Zhenbo; Lu, Weiqing and Cao, Yuran: "A Solution to Co-occurence Bias: Attributes Disentanglement via Mutual Information Minimization for Pedestrian Attribute Recognition". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (California). California: International Joint Conferences on Artificial Intelligence Organization, 2023. DOI: 10.24963/ijcai.2023/203 (cit. on pp. 30, 190).

[Zhu13] Zhu, Jianqing; Liao, Shengcai; Lei, Zhen; Yi, Dong and Li, Stan Z.: "Pedestrian Attribute Classification in Surveillance: Database and Evaluation". In: 2013 IEEE International Conference on Computer Vision Workshops (2013). IEEE, 2013. DOI: 10.1109/ICCVW.2013.51 (cit. on pp. 53, 54).

[Zhu15] Zhu, Jianqing; Liao, Shengcai; Yi, Dong; Lei, Zhen and Li, Stan Z.: "Multi-label CNN based pedestrian attribute learning for soft biometrics". In: *2015 International Conference on Biometrics (ICB)*. 2015 International Conference on Biometrics (ICB) (2015). IEEE, 2015. DOI: 10.1109/ICB.2015.7139070 (cit. on pp. 25, 31).

[Zhu21] Zhu, Aichun; Wang, Zijie; Li, Yifeng; Wan, Xili; Jin, Jing; Wang, Tian; Hu, Fangqiang and Hua, Gang: "DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval". In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21: ACM Multimedia Conference (Virtual Event China). Ed. by Shen, Heng Tao; Zhuang, Yueting; Smith, John R.; Yang, Yang; Cesar, Pablo; Metze, Florian and Prabhakaran, Balakrishnan. New York, NY, USA: ACM, 2021, pp. 209–217. doi: 10.1145/3474085.3475369 (cit. on p. 28).

[Zhu23a] Zhu, Jianqing; Liu, Liu; Zhan, Yibing; Zhu, Xiaobin; Zeng, Huanqiang and Tao, Dacheng: "Attribute-Image Person Re-identification via Modal-Consistent Metric Learning". In: *International Journal of Computer Vision* 131.11 (2023). PII: 1841, pp. 2959–2976. doi: 10.1007/s11263-023-01841-7 (cit. on pp. 38, 46, 191).

[Zhu23b] Zhu, Jun; Jin, Jiandong; Yang, Zihan; Wu, Xiaohao and Wang, Xiao: "Learning CLIP Guided Visual-Text Fusion Transformer for Video-based Pedestrian Attribute Recognition". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Vancouver, BC, Canada). IEEE, 6/17/2023 - 6/24/2023, pp. 2626–2629. doi: 10.1109/CVPRW59228.2023.00261 (cit. on p. 36).

Own Publications

[1] Schumann, Arne; Specker, Andreas and Beyerer, Jürgen: "Attribute-based Person Retrieval and Search in Video Sequences". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (Auckland, New Zealand). IEEE, 2018, pp. 1–6. DOI: 10.1109/AVSS.2018.8639114.

[2] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "An Interactive Framework for Cross-modal Attribute-based Person Retrieval". In: *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, Taipei, Taiwan, 18-21 Sept. 2019*. Institute of Electrical and Electronics Engineers (IEEE), 2019, p. 8909832. DOI: 10.1109/AVSS.2019.8909832.

[3] Eckstein, Viktor; Schumann, Arne and Specker, Andreas: "Large Scale Vehicle Re-Identification by Knowledge Transfer From Simulated Data and Temporal Attention". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2020, Virtual, Online, United States, 14 - 19 June 2020*. Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 2626–2631. DOI: 10.1109/CVPRW50498.2020.00316.

[4] Köhl, Philipp; Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "The MTA Dataset for Multi-Target Multi-Camera Pedestrian Tracking by Weighted Distance Aggregation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,*

CVPRW 2020, Virtual, Online, United States, 14 - 19 June 2020. Institute of Electrical and Electronics Engineers (IEEE), 2020. DOI: 10.1109/CVPRW50498.2020.00529.

[5] Specker, Andreas: "A Realistic Predictor for Pedestrian Attribute Recognition". In: *Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Hrsg.: J. Beyerer; T. Zander*. Vol. 45. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2020, pp. 95–112.

[6] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "A multitask model for person re-identification and attribute recognition using semantic regions". In: *Artificial Intelligence and Machine Learning in Defense Applications II. Ed.: J. Dijk*. Vol. 11543. Proceedings of SPIE. SPIE, 2020, p. 115430I. DOI: 10.1117/12.2573981.

[7] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "An Evaluation Of Design Choices For Pedestrian Attribute Recognition In Video". In: *2020 IEEE International Conference on Image Processing. Proceedings : September 25-28, 2020 : virtual conference, Abu Dhabi, United Arab Emirates*. 2020 IEEE International Conference on Image Processing (ICIP) (Abu Dhabi, United Arab Emirates). Institute of Electrical and Electronics Engineers and IEEE Signal Processing Society. Piscataway, NJ: IEEE, 2020, pp. 2331–2335. DOI: 10.1109/ICIP40778.2020.9191264.

[8] Florin, Lucas; Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "Hardness Prediction for More Reliable Attribute-based Person Re-identification". In: *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 418–424. DOI: 10.1109/MIPR51284.2021.00077.

[9] Moritz, Lennart; Specker, Andreas and Schumann, Arne: "A study of person re-identification design characteristics for aerial data".

In: *Pattern Recognition and Tracking XXXII: SPIE DEFENSE + COMMERCIAL SENSING | 12-17 APRIL 2021. Ed.: M. S. Alam.* Vol. 11735. Proceedings of SPIE. Society of Photo-optical Instrumentation Engineers (SPIE), 2021, Art.–Nr.: 117350P. DOI: 10.1117/12.2587946.

[10] Sommer, Lars W.; Specker, Andreas and Schumann, Arne: "Deep learning based person search in aerial imagery". In: *SPIE DEFENSE + COMMERCIAL SENSING | 12-17 APRIL 2021 - Automatic Target Recognition XXXI. Ed.: R. I. Hammoud.* Vol. 11729. Proceedings of SPIE. Society of Photo-optical Instrumentation Engineers (SPIE), 2021, Art.–Nr.: 117290O. DOI: 10.1117/12.2588179.

[11] Specker, Andreas: "A Step Towards Explainable Person Re-identification Rankings". 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. 107–121.

[12] Specker, Andreas and Beyerer, Jürgen: "Improving Attribute-Based Person Retrieval By Using A Calibrated, Weighted, And Distribution-Based Distance Metric". In: *2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, Anchorage, AK, USA, 19-22 Sept. 2021.* Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 2378–2382. DOI: 10.1109/ICIP42928.2021.9506096.

[13] Specker, Andreas; Moritz, Lennart and Sommer, Lars W.: "Deep learning-based video analysis pipeline for person detection and re-identification in aerial imagery". In: *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies V.* Vol. 11869. Proceedings of SPIE. SPIE, 2021, 118690K. DOI: 10.1117/12.2600109.

[14] Specker, Andreas; Stadler, Daniel; Florin, Lucas and Beyerer, Jürgen: "An Occlusion-Aware Multi-Target Multi-Camera Tracking System". In: *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR) Workshops. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 4168–4177. DOI: 10.1109/CVPRW53098.2021.00471.

[15] Cormier, Mickael; Clepe, Aris; Specker, Andreas and Beyerer, Jürgen: "Where are we with Human Pose Estimation in Real-World Surveillance?" In: *2022 IEEE Winter Conference on Applications of Computer Vision Workshops. WACVW 2022 : 4-8 January 2022, Waikoloa, Hawaii*. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2022, pp. 591–601. DOI: 10.1109/WACVW54805.2022.00065.

[16] Specker, Andreas: "A Transformer based Multi task Model for Attribute based Person Retrieval". In: *Proceedings of the 2021 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*. Vol. 54. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2022, pp. 139–152.

[17] Specker, Andreas and Beyerer, Jürgen: "Toward Accurate Online Multi-target Multi-camera Tracking in Real-time". In: *30th European Signal Processing Conference (EUSIPCO 2022). Proceedings : 29 August-2 September 2022, Belgrade, Serbia*. 2022 30th European Signal Processing Conference (EUSIPCO) (Belgrade, Serbia). Ed. by Routtenberg, Tirza and Tadić, Predrag. European Association for Signal Processing. Piscataway, NJ: IEEE, 2022, pp. 533–537. DOI: 10.23919/EUSIPCO55093.2022.9909670.

[18] Specker, Andreas; Florin, Lucas; Cormier, Mickael and Beyerer, Jürgen: "Improving Multi-Target Multi-Camera Tracking by Track Refinement and Completion". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

Workshops (CVPRW) (New Orleans, LA, USA). IEEE, 2022, pp. 3198–3208. DOI: 10.1109/CVPRW56347.2022.00361.

[19] Specker, Andreas; Moritz, Lennart; Cormier, Mickael and Beyerer, Jürgen: "Fast and Lightweight Online Person Search for Large-Scale Surveillance Systems". In: *2022 IEEE Winter Conference on Applications of Computer Vision Workshops. WACVW 2022 : 4-8 January 2022, Waikoloa, Hawaii*. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers. Piscataway, NJ: IEEE, 2022, pp. 570–580. DOI: 10.1109/WACVW54805.2022.00063.

[20] Cormier, Mickael; Specker, Andreas; Jacques, Julio C. S.; Florin, Lucas; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge: Pedestrian Attribute Recognition and Attribute-based Person Retrieval - Dataset, Design, and Results". In: *2023 IEEE Winter Conference on Applications of Computer Vision workshops. WACVW 2023 : proceedings : 3-7 January 2023, Waikoloa, Hawaii*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers and Computer Vision Foundation. Piscataway, NJ: IEEE, 2023, pp. 166–175. DOI: 10.1109/WACVW58289.2023.00022.

[21] Specker, Andreas; Cormier, Mickael and Beyerer, Jürgen: "UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval". In: *2023 IEEE Winter Conference on Applications of Computer Vision. 3-7 January 2023, Waikoloa, Hawaii : proceedings*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (Waikoloa, HI, USA). Ed. by Berg, Tamara. IEEE Computer Society and Computer Vision Foundation. Piscataway, NJ: IEEE, 2023, pp. 981–990. DOI: 10.1109/WACV56688.2023.00104.

[22] Specker, Andreas and Beyerer, Jürgen: "Balanced Pedestrian Attribute Recognition for Improved Attribute-based Person Retrieval". In: *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*. 2023 IEEE 13th International Conference on Pattern

Recognition Systems (ICPRS) (Guayaquil, Ecuador). IEEE, 7/4/2023 - 7/7/2023, pp. 1–7. DOI: 10.1109/ICPRS58416.2023.10178990.

[23] Specker, Andreas and Beyerer, Jürgen: "ReidTrack: Reid-only Multi-target Multi-camera Tracking". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Vancouver, BC, Canada). IEEE, 6/17/2023 - 6/24/2023, pp. 5442–5452. DOI: 10.1109/CVPRW59228.2023.00575.

List of Figures

1.1 Surveillance use cases . . . 3
1.2 Attribute-based person retrieval ranking . . . 5
1.3 Low-resolution imagery . . . 7
1.4 Different viewpoints . . . 7
1.5 Illumination . . . 8
1.6 Indoor and outdoor scenes . . . 9
1.7 Occlusions . . . 10
1.8 Intra-class variance . . . 10
1.9 Attribute distributions . . . 11

2.1 Taxonomy of soft biometrics . . . 19
2.2 Types of textual queries . . . 25
2.3 Approaches for person retrieval using discrete attribute queries . . . 26
2.4 Comparison of normalization layers . . . 40

3.1 Concept overview . . . 47
3.2 Examples for indeterminable attributes . . . 50

4.1 Example images from the PETA dataset . . . 55
4.2 Example images from the Market-1501 dataset . . . 56
4.3 Example images from the PA-100K dataset . . . 57
4.4 Example images from the RAPv2 dataset . . . 58
4.5 Number of samples per attribute in the UPAR dataset . . . 63

5.1 Baseline architecture . . . 77

6.1 Comparison of backbones architectures . . . 87

6.2 Impact of SWA on the training process . . . 90
6.3 Comparison of SWA methods . . . 92
6.4 Influence of α_{FL} on the retrieval performance . . . 96
6.5 PARNorm . . . 117
6.6 Qualitative evaluation of the PARNorm module . . . 122
6.7 Example of a person track . . . 124
6.8 Temporal pooling approaches . . . 126
6.9 Tradeoff between inference time and retrieval accuracy . . . 132

7.1 Example images depicting challenging factors . . . 134
7.2 Architecture of the HP model . . . 136
7.3 Easy and difficult samples for gender . . . 141
7.4 Easy and difficult samples for jacket . . . 142
7.5 Misclassified samples with contradictory statements . . . 143
7.6 Qualitative comparison of baseline and HP results . . . 150
7.7 Distribution of attribute classifier outputs . . . 153
7.8 Calibration curves for selected attributes . . . 156
7.9 Comparison of reliability calibration methods . . . 163

8.1 Inference times for the proposed model . . . 178
8.2 First qualitative specialization result . . . 180
8.3 Second qualitative specialization result . . . 182
8.4 Third qualitative specialization result . . . 183
8.5 First qualitative generalization result . . . 184
8.6 Second qualitative generalization result . . . 186
8.7 Analysis of error cases . . . 188

9.1 MTA dataset . . . 199
9.2 WDA tracker . . . 202
9.3 Real-world person retrieval system . . . 205

List of Tables

4.1 Overview of PAR datasets . . . 54
4.2 UPAR attribute annotations . . . 61
4.3 Evaluation statistics . . . 72
4.4 UPAR splits . . . 72
4.5 UPAR split statistics . . . 73

6.1 Binary vs. multi-class attributes . . . 85
6.2 Specialization results for different backbones . . . 88
6.3 Generalization results for different backbones . . . 89
6.4 Specialization results for SWA techniques . . . 93
6.5 Generalization results for SWA techniques . . . 94
6.6 Comparison of loss functions for specialization . . . 97
6.7 Comparison of loss functions for generalization . . . 98
6.8 Specialization results for varying batch sizes . . . 100
6.9 Impact of different batch sizes on certain attributes . . . 102
6.10 Generalization results for varying batch sizes . . . 103
6.11 Impact of dropout on the specialization results . . . 105
6.12 Impact of dropout on the generalization results . . . 106
6.13 Comparison of optimizers for the case of specialization . . . 107
6.14 Comparison of optimizers for the case of generalization . . . 108
6.15 Impact of label smoothing on the specialization results . . . 110
6.16 Impact of label smoothing on the generalization results . . . 112
6.17 Summary of PAR model optimization . . . 114
6.18 Summary of PAR model optimization for the UPAR dataset . . . 115
6.19 PARNorm ablation study . . . 120
6.20 Specialization results for the PARNorm module . . . 121

6.21 Generalization results for the PARNorm module 122
6.22 Comparison of normalization techniques 123
6.23 Comparison of temporal pooling operations 128
6.24 Comparison of video processing methods 129
6.25 Motion classification results 131

7.1 HP ablation study 145
7.2 HP specialization results 147
7.3 HP generalization results 147
7.4 Comparison with self-referential approaches 148
7.5 Specialization results for reliability calibration methods . . . 161
7.6 Generalization results for reliability calibration methods . . . 162
7.7 Supervised vs. unsupervised specialization results 165
7.8 Supervised vs. unsupervised generalization results 166
7.9 Retrieval improvements specialization results 166
7.10 Retrieval improvements generalization results 167

8.1 PARNorm results 170
8.2 Specialization results 172
8.3 Generalization results 173
8.4 UPAR LOOCV split results 175
8.5 UPAR 4FCV split results 176
8.6 State-of-the-art PAR 190
8.7 State-of-the-art attribute-based person retrieval 191
8.8 State-of-the-art UPAR LOOCV 192
8.9 State-of-the-art UPAR 4FCV 193

Acronyms

4FCV	4-Fold Cross-Validation
AP	Average Precision
CMC	Cumulative Matching Characteristics
CNN	Convolutional Neural Network
DBD	Distribution-Based Distance
DoM	Degree of Match
EMA	Exponential Moving Average
F1	F1
FC	Fully-Connected
FPS	Frame Per Second
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GDPR	General Data Protection Regulation
GMP	Global Maximum Pooling

GPU	Graphics Processing Unit
HP	Hardness Predictor
IDF1	Identity F1
LOOCV	Leave-One-Out Cross-Validation
LSTM	Long Short-Term Memory
mA	Mean Accuracy
mADM	Mean Average Degree of Match
mAP	Mean Average Precision
MARS	Motion Analysis and Re-identification Set
MLP	Multi-layer Perceptron
MSE	Mean Squared Error
MTA	Multi-camera Track Auto
MTMCT	Multi-Target Multi-Camera Tracking
PA-100K	Pedestrian Attribute 100K
PAR	Pedestrian Attribute Recognition
PETA	Pedestrian Attribute
PVTv2	Pyramid Vision Transformer v2
R-1	Rank-1 accuracy

RAPv1	Richly Annotated Pedestrian v1
RAPv2	Richly Annotated Pedestrian v2
ResNet	Residual Net
SMA	Simple Moving Average
SVM	Support Vector Machine
SWA	Stochastic Weight Averaging
SWAD	Stochastic Weight Averaging Densely
TFP	Temporal Feature Pooling
TPP	Temporal Prediction Pooling
UPAR	Unified Pedestrian Attribute Recognition
ViT	Vision Transformer
WDA	Weighted Distance Aggregation

Symbols

Acc@k	Accuracy at rank k across multiple queries
α_{EMA}	EMA hyperparameter
α_{FL}	Focal loss hyperparameter
α_{LS}	Label smoothing hyperparameter
B	Batch size
β	Shift parameter
C	Number of channels
$\mathcal{C}$	Classifier function
D	Sample size
$\mathcal{D}$	Dataset
δ_{HP}	Hardness hyperparameter
DoM@k	Degree of match at ranking position k
$\overline{\text{DoM}}$	Mean degree of match across the gallery
DoM_{Norm}@k	Normalized degree of match at ranking position k

ϵ	Small constant
F	Feature dimension
$\mathrm{F1}_{\mathrm{PAR}}$	Instance-based F1 score for pedestrian attribute recognition
f	Function that returns the recognized person attributes for an image
$\mathcal{G}$	Gallery set
G	Gallery size
γ	Scale parameter
GTP	Number of ground truth positives
H	Image Height
H_{f}	Feature map height
$\mathbf{I}$	Image
k	Position in retrieval ranking
L	Number of semantic attributes
$\mathcal{L}$	Loss function
$\mathcal{L}_{\mathrm{CE}}$	Cross-entropy loss function
$\mathcal{L}_{\mathrm{FL}}$	Focal loss function
$\mathcal{L}_{\mathrm{HP}}$	Loss function used to train the hardness predictor

$\mathcal{L}_{\text{PAR}}$	Loss function used to train the PAR classifier
M	Number of images
mA	Label-based mean accuracy
$\mathbf{m}$	Vector of mean squared errors
m	Mean squared error
μ	Mean value
N	Number of negative samples
N	Number of attributes included in a query
$\mathbf{p}$	Predicted attribute confidence vector
P	Number of positive samples
$\text{Prec}_{\text{DoM}}@k$	Precision score calculated based on the degree of match at rank k
$\text{Prec}_{\text{IR}}@k$	Information retrieval precision score at rank k
Prec_{PAR}	Instance-based precision score for pedestrian attribute recognition
p	Predicted attribute confidence value
p^y	Predicted probability for the ground truth value
$\mathbf{q}$	Attribute query
q	Attribute query value

r_{drop}	Positive ratio
Rec_{PAR}	Instance-based recall score for pedestrian attribute recognition
Rel@k	Relevance indicator for information retrieval
$\mathbf{s}$	Hardness vector
s	Hardness value
T	Tracklet length
$\mathbf{t}$	Attribute binarization threshold vector
t_0	Stochastic weight averaging start iteration
t_e	Stochastic weight averaging end iteration
θ	Learnable parameters of the entire model
$\hat{\theta}$	Averaged model parameters
θ_{c}	Learnable parameters of the classifier
θ_{f}	Learnable parameters of the feature extractor
t	Training iteration
TN	Number of true negatives
TP	Number of true positives
W	Image width

w	Attribute-specific loss weight
$\mathbf{w}^{\text{err}}$	Attribute-specific error weight vector
W_{f}	Feature map width
$\mathbf{x}$	Logit vector
x	Logit value
Y	Set of ground truth positive labels
$\mathbf{y}$	Binary attribute label vector
y	Binary attribute label
z	Distribution-based distance value