Pose-Guided Semantic Person Re-Identification in Surveillance Data

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

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

genehmigte

Dissertation

von

Arne Schumann

aus Lüneburg

Tag der mündlichen Prüfung:	05.07.2019
Erster Gutachter:	Prof. Dr.-Ing. Rainer Stiefelhagen
Zweiter Gutachter:	Prof. Dr. Shaogang Gong

Abstract

The rapid increase in available surveillance imagery can be a great asset to law enforcement and security services. However, manual analysis of such large amounts of data becomes ever more infeasible. Intelligent surveillance systems are required to aid human operators in their task. The *re-identification* of persons is a key functionality of such systems. Short-term re-identification of people based on their full-body appearance enables online search tasks, such as localizing the current position of a person of interest, even if their face is not recognizable, due to the commonly low resolution of surveillance footage or obstructing camera views. Person tracking approaches can similarly benefit from support of re-identification methods to bridge gaps between cameras or occlusions between people or by scene elements. By far the most established application for person re-identification is, however, the offline retrieval task. In offline retrieval, large amounts of images or video sequences have to be searched in order to reconstruct the movements of one or more individuals. For this, an automatic approach computes a feature representation of person images and matches these representations in order to establish scores for visual similarities between two person images. The result consists of a list of images ranked according to the similarity to a query image. A successful representation and matching approach must address a number of challenging factors which affect a person's visual appearance. This includes changes in illumination, camera view, body pose, image noise, and heterogeneous sensor characteristics.

The aim of this thesis is to develop full-body person re-identification models with high robustness to these challenges. A focus lies on aspects pertaining to real-world deployment of re-identification systems. This includes the requirement for adaptability to new or changing camera networks, the application of automatic person detectors and their resulting errors, the degree of semantic information which makes the result interpretable to operating personnel, as well as scalability with very large numbers of candidate person images that are to be searched.

To this end, two approaches for different application scenarios are proposed: i) a supervised method that can be used when training data is available to adapt the model to a given target scenario and ii) a re-identificiation framework which can automatically adapt to new scenarios without the need for additional annotation or training data at all. For the supervised approach two methods of explicitly modeling pose-related information into an initial Convolutional Neural Network (CNN) are proposed. Both, full-body pose in the form of body joint key points, as well as coarse view information are included and shown to be complementary. Furthermore, a second CNN model is developed which includes information from previously learned semantic attributes and learns image features, which are complementary to the attribute information. This adds a degree of interpretability to the results and attributes can further help to strengthen robustness to illumination and sensor characteristics. Aspects from both models are combined into a final supervised model, which yields state-of-the-art accuracy on public datasets. The second approach targets a deployment scenario in which no data is available for adaptation to the target scene. The approach learns a set of prototypical subsets in a large amount of diverse training data. These subsets, termed prototype domains, each represent a combination of typical scene and person characteristics. A collection of models is trained, each specialized on one of the domains. When new re-identification queries occur at test time, the model most closely related to the query image is chosen for the task. Thus a query-adaptive approach is achieved which requires no more data than the query image itself. Both approaches are additionally evaluated using more challenging settings pertaining to real-world deployment, such as very large gallery sizes and use of automatic person detectors.

Kurzfassung

Die großen Mengen verfügbarer Bild- und Videodaten aus Kameranetzwerken können eine große Hilfe für Strafverfolgung und Sicherheitsdienste sein. Eine manuelle Auswertung der wachsenden Datenmengen ist jedoch in vielen Fällen bereits heute nicht mehr mit vertretbarem Aufwand zu leisten. Intelligente Systeme zur automatischen Auswertung werden benötigt, um menschliche Analysten zu unterstützen. Die automatische *Wiedererkennung von Personen* über kurze Zeiträume hinweg ist eine zentrale Funktion solcher Systeme. Eine automatische Wiedererkennung anhand von Ganzkörper-Merkmalen unterstützt beispielsweise Aufgaben wie die Echtzeitsuche, bei der die aktuelle Position einer Person innerhalb eines Kameranetzwerkes gefunden werden soll. Insbesondere ist es hierbei auch möglich Personen zu finden, deren Gesicht aufgrund von schlechter Bildqualität oder ungünstigem Kamerawinkel nicht erkennbar ist. Ein weiteres Einsatzszenario ist die offline Suche, während der große Datenmengen prozessiert werden müssen, um anschließend automatisiert nach Personen durchsucht werden zu können. Hierzu berechnen und vergleichen Verfahren zur Wiedererkennung interne Merkmalsrepräsentationen von Personenbildern und erzeugen numerische Werte, welche die visuelle Ähnlichkeit zwischen Personen abbilden. Das Ergebnis der Wiedererkennung ist eine Liste von Personenbildern, die nach der visuellen Ähnlichkeit zu einem Anfragebild sortiert ist. Die Verfahren müssen dabei eine Reihe von Faktoren handhaben, welche die visuelle Erscheinung von Personen im Bild

beeinflussen. Diese beinhalten Beleuchtungsunterschiede, variierende Kamerawinkel, Körperposen, Bildrauschen und Unterschiede zwischen den eingesetzten Kamerasensoren.

Das Ziel dieser Arbeit ist die Entwicklung von Modellen zur Wiedererkennung von Personen, die diese Herausforderungen adressieren. Ein Schwerpunkt ist hierbei die Betrachtung von Anforderungen, die beim Einsatz der Modelle in realen Szenarien entstehen. Beispielsweise müssen sich die Modelle ohne erheblichen manuellen Aufwand an neue und unbekannte Szenarien anpassen können. Auch der Einsatz automatischer Verfahren zur Detektion von Personen und der Fehler, die hierbei entstehen, wird betrachtet, sowie die Interpretierbarkeit der Ergebnisse durch menschliche Benutzer und die Skalierbarkeit mit großen Mengen von Personen.

Zwei Ansätze werden in dieser Arbeit verfolgt, die sich auf unterschiedliche Szenarien spezialisieren. Ein Szenario, indem es anhand von annotierterten Daten möglich ist Verfahren an die vorliegenden Charakteristiken der Daten anzupassen, und ein Szenario, in dem keine solche Daten zur Verfügung stehen.

Um den ersten Fall zu adressieren, werden zunächst zwei separate Modelle vorgeschlagen. Das erste Modell bezieht semantische Attribute in den Prozess der Wiedererkennung mit ein und erhöht somit die Genauigkeit und Interpretierbarkeit der Ergebnisse. Das zweite Modell kombiniert Informationen der Körperpose und des Sichtwinkels der Kamera, um eine Merkmalsrepräsentation zu erlernen, die robust gegenüber diesen Variationen ist. Die Erkenntnisse beider Modelle fließen schließlich in ein kombiniertes Verfahren ein, das hohe Genauigkeit auf mehreren öffentlichen Datensätzen erreicht.

Um das zweite Szenario zu adressieren, wird ein Verfahren bestehend aus einer Menge von Modellen vorgeschlagen. Jedes Modell wird hierbei auf eine bestimmte Ausprägung von typischen visuellen Charakteristiken spezialisiert, welche automatisch anhand einer Datenmenge mit hoher visueller Vielfalt identifiziert werden. Bei einer Suchanfrage in Daten mit unbekannten Charakteristiken wird anhand des Anfragebildes ein geeignetes Modell aus der verfügbaren Menge gewählt. Somit kann eine automatische Anpassung

an die aktuelle Suchanfrage realisiert werden, die keinerlei zusätzliche Trainingsdaten voraussetzt.

Beide Modelle erzielen Ergebnisse, die dem Stand der Forschung entsprechen und werden zusätzlich auf ihre Kooperation mit automatischen Personendetektoren und die Erkennungsleistung bei sehr großen Personenmengen hin untersucht.

Contents

Chapter 1

Introduction

Person re-identification (re-id) is the problem of, given a query image of a specific person, finding further occurrences of that same person in large amounts of image or video data based on the full-body appearance. The most common application of person re-identification is intelligent video surveillance. Typical surveillance camera networks contain non-overlapping camera views, uncalibrated cameras, and changes in the network's topology. Tracing a specific person's path through the camera network thus requires time and a high degree of attention, if attempted manually by security personnel. In most practical applications such manual analysis is not possible. For example, when a suspicious item of left luggage has been detected in a public area, such as a shopping mall, locating the current whereabouts of the person who left the item can be a very time-sensitive issue. A second common usecase is offline data analysis, such as review of Closed Circuit Television (CCTV) footage after a security incident with the goal of following the movements of suspects in the aftermath. Such an offline scenario may not be as time-sensitive but a manual analysis is nevertheless prohibitive due to the enormous amount of video data that is generally available. Automated or semi-automated person re-identification systems are thus a crucial component of modern intelligent surveillance systems. A typical workflow requires the operator to select an initial occurrence of a person of interest as query image and the system

then returns a list of persons from the search database (often termed *gallery*) ranked according to visual similarity to the query image. The operator can then identify correct matches among the top-ranked results or may have the option to provide feedback to the system to improve results in a next iteration [Sch15c].

While surveillance scenarios are the most prominent real-world application and the main focus of the research community, several other applications exist which can be addressed supported by person re-identification. For example, identification of actors can help to provide viewers of TV series or movies with interesting metadata. In such applications face matching is used to perform actual identification and person re-identification can be applied to connect the dots when faces are not visible [Bäu13]. A similar combination of approaches can be applied for human-computer-interaction in smart home environments or for patient monitoring in care facilities. The main focus of this thesis, however, lies on person re-identification in surveillance settings, due to the large amount of available public datasets in this area.

Historically, matching of person identities across different images was first addressed through face recognition in 1977 [Kan77]. In contrast, one of the first works on full-body based person re-identification was proposed in 2005 [Zaj05]. There are several core differences between face-based and full-body person re-identification which make them suitable for different types of applications but they also complement each other. Faces are often more discriminative whereas full-body appearance is mostly determined by clothing, which may appear very similar among different individuals, due to fashion trends or work uniforms. On the other hand, full-body person re-identification is applicable when faces are not visible due to view angles or occlusions caused, for example, by scarfs, hair, or sunglasses. Furthermore and most importantly, full-body person re-identification is applicable at much lower resolutions when faces are no longer recognizable. This is often the case in surveillance scenarios where low resolution cameras are prevalent due to cost considerations. Finally, full-body re-identification is less invasive to personal privacy, because the internally used representations generally do not allow for an actual identification of individuals. This aspect is of particular

interest in countries with stronger restrictions to preserve privacy and data protection, such as Germany.

Person re-identification approaches can rely on several types of image cues and features. Arguably the most relevant type of information for re-id in RGB images is color information. This is often combined with texture or contour cues. Soft-biometric features or attributes (e.g. gender, hair color, clothing descriptors) can add an additional level of semantics to the re-id task. Furthermore, the automatic recognition of such attributes enables additional applications, such as re-id based purely on a witness description. When video is available, gait information may be leveraged to support re-id and 3D sensor data allows for integration of physical measurements, such as body type.

1.1 Challenges

Person re-identification approaches for camera networks must overcome several challenges. These can be grouped into two main categories. Challenges resulting from the scene or environment in which re-id is to be carried out and challenges arising during the image capturing process in the sensors deployed in the camera network.

Scene Challenges

- **Illumination** can vary strongly in different areas of the camera network. This is particularly the case when indoor and outdoor areas are part of the network. Low illumination can result in a reduced signal-to-noise ratio while strong illuminations will result in strong shadows. Both extremes of illumination will decrease contrast and alter color intensities.
- **Partial Occlusions** between people or by scene elements, such as low walls, poles, or trashcans can hide important aspects of the visual appearance of a person. In the worst case, appearance aspects of the occluding person may distort the feature representation of the occluded person and result in faulty matchings.

- **Different Viewpoints** of persons relative to the camera can have the most significant impacts on visual appearance. For example, the appearance of a person wearing a t-shirt and backpack will in frontal view be dominated by the t-shirt while a view from the back will rather be influenced by the appearance of the backpack. If a re-id approach is unaware of differences in viewpoint while attempting to compare person images, unwanted mismatches can occur.
- **Different Body Poses** of a person can alter the location of important details within a person image. This is particularly the case for legs of pedestrians which are generally moving. But also upper-body pose, such as an arm crossing the torso or hanging to the side can affect the overall appearance and location of relevant image details.
- **Small Details** are often of key relevance for successful re-id . Due to fashion trends or work uniforms, the region in the image that is actually relevant for successful discrimination between two similar looking people may be very small.

Sensor Challenges

- **Resolution** of images in typical surveillance networks is often low, due to use of cost-efficient cameras. Furthermore, due to the placement of surveillance cameras (at ceiling level, looking down hallways, streets, or places), significant differences in resolution of person images between foreground and background may occur. A typical resolution of person images that allows for re-id at realistic and practical distances is 64×128 pixels.
- **Varying Sensor Characteristics**, such as differing color sensitivity resulting in color tints, can lead to differences in images even under uniform illumination. Such effects may occur in heterogeneous camera systems, for example when broken cameras are replaced by different models over time.
- **Image Noise** resulting from low illumination, quantum noise, measurement noise or quantization noise during the image capturing

process may affect and distort important details of the visual appearance.

- **Blur** caused by quick motion in the image, too long exposure times, faulty focus settings, or smudges or dirt on the lense can similarly destroy crucial details in the image.
- **Compression Artifacts** caused by image compression methods can reduce image quality and level of detail. In order to transmit and store the large amount of data acquired in surveillance camera networks, compression is often unavoidable due to limitations in bandwidth, write-speed, and overall storage capacity.

A number of these challenges are illustrated in Figure 1.1. To varying degree, all of these challenges affect the visual appearance of persons and contribute to the core issue that needs to be addressed by any re-id method: In image space, a high degree of intra-class variance can exist while simultaneously the degree of inter-class variance may be very low. Or, in other words, the difference in visual appearance of the same person across multiple images may be high, due to variation in view, illumination, and image noise, while the difference in visual appearance of different people may be low, due to similar clothing. The task of a re-id approach is then to invert this relationship and thus allow for a clear separation between different individuals while maintaining high similarities for images of the same individual. Generally speaking, this can be achieved in one of two main ways. The person images can be transformed into a feature space that provides the desired properties when used with a standard distance metric to compare representations. Such feature transformations can be handcrafted through careful selection and encoding of relevant image information or learned from large amounts of data using machine learning methods. Alternatively, a suitable image feature can be used with a well-engineered or learned distance metric which provides the desired variance properties. The methods described in this work mainly focus on the former alternative.

Figure 1.1: Person images from the VIPeR [Gra08], Market-1501 [Zhe17b], and Duke MTMC-reID [Zhe15b] datasets. The images illustrate visual variations caused by image quality, viewpoint, occlusions, and pose.

1.2 Recent Developments

The initial accuracy of re-id approaches was of limited use for practical surveillance applications. However, over the last years research interest and matching accuracy have increased significantly. This is illustrated, for example, by the rise in state-of-the-art matching accuracy on the popular Market-1501 dataset [Zhe15a]. At publication of the dataset in 2015 the rank-1 accuracy was 44.4%, i.e. 44.4% chance that the first rank in the result list would contain a correct match to the query. In contrast, the accuracy achieved by methods proposed in this work is 92.1%. Similarly high values are achieved by several recent approaches. This high accuracy has significantly increased the practical use of re-id and enabled the development and deployment of real-world systems. However, the prospect of real-world application brings with it a new set of challenges.

Challenges of real-world application

- **Generalization or Adaptability** - In real-world applications, re-id systems are deployed in new scenarios which result in data different to that on which the underlying models were trained. The annotation of new data from the target scenario for retraining or adaptation of the models is often prohibitively expensive. A core challenge of such models is thus to achieve either good generalization to a large range of possible target scenarios or to adapt to target scenarios without the need for annotated data.

- **Scalability** - Research datasets often feature comparatively small galleries in which query persons must be found. Typical are thousands to few ten thousands of gallery images. In many real-world scenarios the number of people passing through the camera network in any given day is much larger. Judging a model's accuracy for large gallery sizes of hundreds of thousands of images gives important hints for their applicability in real-world scenarios.
- **Interpretability** - Trust is an important aspect for the application and acceptance of any intelligent system. If operators do not trust a system, they will avoid relying on it. A key component to establishing such trust with operating personnel is to provide an understanding of why the system arrives at certain results, especially in case of faulty or unexpected results. Providing information for decisions that is interpretable by non-expert users is thus an important feature for real-world application.
- **Person Detection & Search** - Many research datasets provide manually pre-cropped person images. In real-world applications, however, re-id must be used in combination with an automatic person detector in order to be efficient. Such a detector will generate errors, such as mis-aligned bounding boxes, partial detections, or false-positive detections which the re-id system should be able to handle. The problem setting where query person images must be found in whole camera images and both person detection and re-id have to be carried out consecutively or jointly is generally referred to as person search.

1.3 Contributions and Outline

The contributions of this thesis focus on developing approaches for person re-identification that successfully address the previously discussed challenges. To that end, two different models are proposed. The central element of the thesis is a model which incorporates information from semantic attributes, as

well as pose information into the re-id representation. The model is trained in a supervised manner and evaluated in within-dataset settings. The second proposed model is specifically designed for deployment to new scenarios without the need for additional training data and is evaluated in a cross-dataset setting. Both models are evaluated on public datasets and with regard to real-world application. After a review of the literature related to the proposed approaches in Chapter 2, the core motivations and concepts of the developed models are outlined in Chapter 3.

The remainder of the thesis describes the proposed approaches based on the following summarized contributions.

In CHAPTER 4: SEMANTIC ATTRIBUTES, a method for recognition of semantic person attributes is formulated.

- The approach uses recent deep learning models and recognizes many attributes with a single model. The proposed model combines local and global person information and incorporates pose information in order to better localize and focus on small image details which are often crucial for successful recognition of more subtle attributes. The approach achieves state-of-the-art accuracy on several public datasets. Uniquely, the proposed approach demonstrates that pose information can be a useful basis for attention mechanisms in the context of attribute recognition. A precursor to this approach has been published in [Sar17b].

In CHAPTER 5: ATTRIBUTE AND POSE SENSITIVE RE-IDENTIFICATION, two novel re-id models are proposed, which incorporate attribute and pose information, respectively.

- In order to incorporate the attribute cues generated by the model described in Chapter 4, a novel variant of the triplet loss is proposed, which allows for learning of re-id features that are complementary to the information contained within attribute cues. The resulting combination of attributes and complementary features is shown to result in improved accuracy compared to established feature fusion methods [Sch17b].

- Viewpoint variation as well as unusual body poses are addressed by a model which incorporates full-body pose information based on 14 body joint keypoints, as well as a view classification stage which allows elements of the model to focus on one of three distinct views. Both aspects are shown to have complementary benefits and result in much improved matching accuracies [Sar18b].
- Finally, a joint re-id model is proposed, which incorporates attribute and pose information. The model is shown to achieve a further increase in re-id accuracy and is evaluated under several real-world aspects, including very large person databases, application with automatic person detectors, and video-based re-id. In combination with re-ranking methods, this model achieves state-of-the-art accuracies on several public datasets.

In CHAPTER 6: DOMAIN PROTOTYPE LEARNING, the real-world challenge of deploying a model in scenarios with new and unknown characteristics and biases is addressed. To that end

- a strategy is proposed to identify groupings with similar scene or person characteristics in large amounts of highly diverse person imagery. These groupings can be identified as prototypical re-id domains and each combines a certain set of visual and salient person or scene attributes.
- The identified prototype domains are then used for development of a flexible re-id framework which can adapt to new scenarios and even individual queries through efficient selection of relevant domain-specific re-id models for the best fitting prototype domain [Sch17a]. The approach is the first to achieve a level of adaptation to new data without requiring either supervised or unsupervised re-training.

A summary of the main findings, as well as suggestions for promising future research directions are given in Chapter 7.

Chapter 2

Related Work

2.1 Person Re-Identification

The task of person re-identification has received a significant growth in attention from the research community, as well as industry in recent years. This section will first review the large corpus of existing literature, most of which focuses on methods based on CNNs. An overview of landmark and recent datasets will be given, which offer a variety of different evaluation settings for re-id methods. Finally, the established metrics to quantitatively compare results of different re-id methods will be reviewed.

2.1.1 Existing Methods

The current re-id literature can be grouped into several categories. This review will first discuss methods based on conventional, hand-crafted and in some cases learned person feature descriptors. Then, metric learning techniques will be discussed, followed by a review of deep learning methods, which make up the majority of ongoing research directions. The inclusion of additional information, such as semantic attributes and pose information,

as employed in the approaches proposed in this thesis, will be addressed separately. Finally, recent methods for re-ranking, a popular post-processing step, will be reviewed.

2.1.1.1 Person Descriptors

One of the earliest approaches to re-id was the manual design of suitable feature descriptors, which encode information relevant to robustly match persons across cameras. Given a well crafted feature, matching is then usually carried out through a standard distance measure, such as euclidean or cosine distance. Hand-crafted person feature descriptors focus primarily on color and in some instances on gradient or texture information.

An early descriptor by Gheissari et al. [Ghe06] applies a segmentation approach to detect temporally stable regions. In such regions, color and edge histograms are then computed. A more complex feature including 8 color channels (RGB, HS, and YCbCr) and 21 texture filters is proposed in [Gra08]. The authors also propose to split the person image into horizontal stripes to better maintain spatial information. The technique of horizontal striping is used in a number of re-id approaches, including recent ones based on deep learning. A further instance of a stripe-based descriptor is proposed by Mignon et al. [Mig12] and contains several color channels, as well as histograms of local binary patterns (LBP). One of the most well-known handcrafted re-id descriptors is the Symmetry Driven Accumulation of Local Features (SDALF) descriptor by Farenza et al. [Far10]. The SDALF feature is based on a weighted color histogram (WH), maximally stable color regions (MSCR), and recurrent high-structured patches (RHSP), which is a texture descriptor. An 11-dim color names descriptor is proposed in [Van09]. The descriptor is computed on local patches and combined into a single vector through a Bag-of-Words (BoW) approach. Pedagadi et al. [Ped13] extract color histograms and color moments from the HSV and YUV spaces and apply Principal Component Analysis (PCA) for dimensionality reduction. The idea of horizontal striping to include local information is further expanded in [Zha13] where small local patches are extracted and LAB color histograms as well as SIFT

features are computed. An adjacency search is carried out to find local patch matches within horizontal directions [She15]. Another landmark re-id feature is proposed by Liao et al. in [Lia15a]. The Local Maximal Occurrence feature descriptor (LOMO) consists of color and SILTP histograms. Max pooling is applied to group feature elements into bins across horizontal stripes. A pyramid approach is applied to include information at multiple scales. With the onset of deep learning methods for the re-id task, development of further hand-crafted descriptors has received little attention in recent years.

2.1.1.2 Metric Learning

While hand-crafted descriptors often rely on standard distance metrics for matching, these might not always be the optimal choice. Particularly for high-dimensional features, a better adapted distance metric may lead to significantly increased matching accuracies. The task of metric learning aims at learning a suitable distance metric from data. Usually this requires pairs of person images showing either the same person or different persons. The aim of the metric learning process is then to modify the metric in such a way that descriptor vectors for images with the same person ID result in smaller distances and descriptor vectors for images of different persons result in larger distances. The most frequently used metric learning methods are based on the Mahalanobis distance. The Mahalanobis distance is a generalization of the euclidean distance and can be written as follows

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \tag{2.1}$$

for two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ and a positive semidefinite matrix $\mathbf{M}$.

An early metric learning method for re-id is proposed by Weinberger et al. [Wei09]. The large-margin nearest-neighbor learning (LMNN) defines a radius for positive person ID matches and penalizes negative matches that fall within less than a margin of this radius. The method is later improved by Davis et al. [Dav07] to increase robustness against overfitting. The most widely used metric learning method for re-id is the KISSME approach [Koe12].

It is based on Mahalanobis distance metric learning and formulates a likelihood ratio test, whether a pair of feature vectors is similar or not. The authors show that the Mahalanobis distance metric can be derived from the log-likelihood ratio test. PCA is applied to the data points to eliminate dimension correlations.

A reduced computational cost was achieved by Hirzer et al. [Hir12] through relaxing the positivity constraint. In [Lia15b], Liao et al. apply different weighting of positive and negative samples. Yang et al. [Yan16c] take into account differences and similarities between image pairs and show that the covariance matrices of negative pairs can be computed from covariance matrices of positive pairs. This process allows the metric learning approach to scale to larger datasets. In combination with the LOMO feature descriptor, Liao et al. [Lia15a] propose to learn a projection to a low-dimensional subspace with cross-view data, as well as learning a metric on that subspace. The approach is similar to linear discriminant analysis (LDA) [Mik99].

2.1.1.3 Deep Learning

Deep learning-based approaches to re-id can be split into two main types of network models: *classification models*, which apply a classification loss function, and *ranking models*, which rely on a ranking loss that compares images of two or more persons.

Classification Models

Classification-based re-id models are the first type of deep learning models used (see Figure 2.1). The general approach to using classification models for re-id is to train them for person ID classification, i.e. using a final layer in the network of dimension equal to the number of person IDs in the training data. Standard methods and losses from the wider image classification literature can then be employed to train the model. For matching of previously unseen persons at test time, the final ID layer is discarded and the prior layer is used as a feature vector. Person images are then ranked according to this feature

vector and a standard distance measure, such as euclidean or cosine distance.

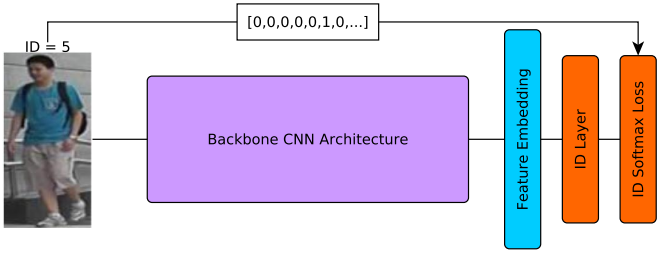

Figure 2.1: Standard classification model for re-id. The model is trained to classify numerical person IDs represented as a binary vector with a one-hot encoding. The last layer before the ID classification typically serves as a feature embedding for re-id and can have an arbitrary dimension. Loss and ID layer are not used at test time.

A key requirement for use of classification models in any task is the availability of sufficient training samples for each class. In re-id early datasets often only contain two images per person. Thus, early works on re-id using deep learning do not use direct classification of person IDs but rather rely on architectures with siamese elements, which require two images to be processed jointly (see Figure 2.2). The classification layer is then set up to learn whether these two images depict the same person or different persons.

In one of the first of such approaches, Li et al. [Li14] describe a filter pairing network to model translation, occlusion and background clutter inside the network's architecture. The final loss function is a softmax cross-entropy loss with only two classes, i.e. same and different. This strategy of relying on pairs of images during training allows for the creation of many more samples per class. Ahmend et al. [Ahm15] introduce a neighborhood matching layer as fusion component, which computes distances between the features resulting from both branches by taking local neighborhoods into account and relying on the smallest found distance. This introduces a degree of robustness

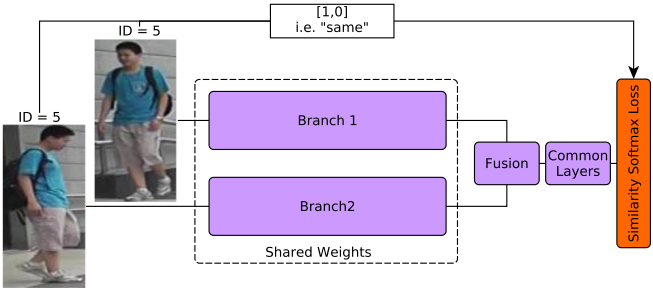

Figure 2.2: Early similarity classification model for re-id. The model requires image pairs. Both images are processed in two siamese branches with usually shared weights. The resulting features are merged by a fusion component and finally, the classification loss learns the binary decision whether the two images depict the same person or not.

to translation and pose change. This neighborhood matching layer is also applied by Wu et al. [Wu16a] to train an end-to-end re-id net which directly outputs a (dis-)similarity decision without relying on a separate distance function. Finally, Varior et al. [Var16a] use a similarity classifier in combination with gating function after each convolutional block. While these methods can be trained on data where the amount of images per ID is limited, they possess the strong drawback that for a comparison between a query image and a gallery database, the query image has to be passed through the network with every gallery image, because the matching and (dis-)similarity decision happens within the layers of the network. This requirement leads to a significant reduction in runtime, as no features can be pre-computed.

With increasing size of datasets and specifically an increase of the number of images per person classification losses can now be used directly for ID classification. Wu et al. [Wu16b] propose a feature fusion deep neural network in which they combine hand-crafted features with learned CNN features. Therefore, the hand-crafted feature vectors are transformed through a fully-connected layer and merged with the fully-connected feature layer of the main CNN branch. A softmax classification loss is used to train the joint network. A classification approach involving training across several datasets and

a large number of IDs is described by Xiao et al. [Xia16a]. Dataset-specific dropout is used on the fully-connected layers near the back of the network in order to allow certain neurons in the layers to adapt specifically to certain datasets. The network is trained with a joint classification loss across all datasets but different elements of the final layers are activated only depending on the dataset each image belongs to. Su et al. [Su17a] combine a global and several local network paths by using a joint classification loss, as well as two separate classification losses on each of the branches. Multiple image scales for cross-resolution re-id are used in [Che17]. Each branch has its own classification loss and a combined loss fuses the information of separate branches. Multiple losses are also used by Li et al. [Li17e]. Local information is extracted in a separate network branch by horizontal splitting of a feature map, learning specialized local features in separate branches and combining these for a final local softmax cross-entropy loss. In addition to this a standard global softmax cross-entropy loss is used. Similarly, Sun et al. [Sun18] apply horizontal striping within a feature layer of the network but attach a separate classification loss to each resulting stripe of the feature map. While most other methods rely on initialization through ImageNet-pretrained models, Li et al. [Li18b] propose an attention-based network architecture that can be trained from scratch and thus requires much fewer parameters. Attention-based approaches rely on learning a weighting of feature maps with a limited weight budget, usually induced through a softmax operator. Since not all weights can be 1.0 under this limitation, the network is forced to assign higher weight, i.e. pay attention, to the most relevant features.

Ranking Models

Similar to the early approach on classification models, ranking CNN models require two or more images as input (see Figure 2.3). However, the loss function of the network does not aim to perform a similarity classification, but rather computes a distance value and optimizes the distances between images showing the same person and images showing different persons. Thus, these loss functions directly optimize a distance-based ranking between two or more images. Importantly, such approaches only perform the matching

operations in the loss function and do not contain any earlier layers, which fuse information from the different branches. Thus, after training, the network can be used to extract feature vectors of single images which can be matched quickly at test time. Networks using such ranking losses based on two and three images are typically referred to as siamese [Rad16] or triplet [Sch15b] models.

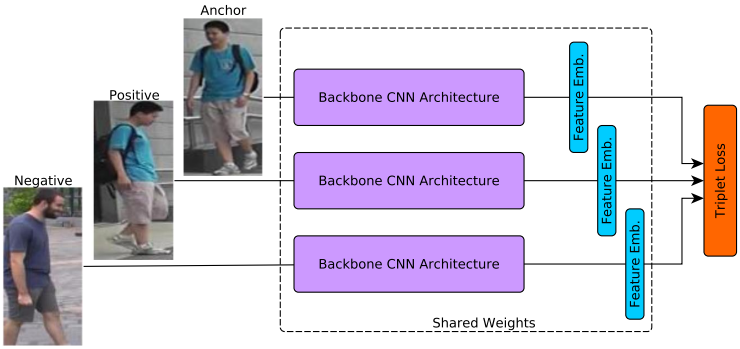

Figure 2.3: Triplet loss ranking model for re-id. An anchor image, a positive match to the anchor, and a negative match to the anchor are provided at once. The loss function optimizes the feature embedding so that the distance from positive to anchor is smaller than the distance from negative to anchor. The weights of each branch are shared and the structure of the model at test time is the same as that of a classification model.

The first work to rely on a distance computation loss is proposed by Yi et al. [Yi14]. Person images are split into three body regions and three siamese networks are used to learn features for each of them. The part features are combined into a final feature through a fully connected layer and distances are computed based on cosine distance. Liu et al. [Liu17a] use a soft attention mechanism in a siamese network to focus on local parts for matching image pairs.

Ding et al. [Din15] apply a triplet loss function involving three images. An anchor image, a positive match to the anchor, and a negative match are passed

through the network and the loss function aims to reduce the euclidean distance between the anchor and the positive match while increasing the euclidean distance between the anchor and the negative match. An improved triplet loss function, which additionally emphasizes small absolute distances for matching image pairs is proposed by Cheng et al. [Che16]. An extensive study of the triplet loss is provided in [Her17]. Hermans et al. propose to generate image triplets automatically for each batch just prior to computation of the loss. Thus, the approach does not rely on a network structure with multiple branches. A part-based method is introduced in [Zha17c]. The method splits person images into multiple body parts and applies a triplet loss function separately for each part. Yu et al. [Yu18] describe a generalization of the triplet loss and compare not just single image pairs but instead an anchor image to sets of positive and negative matches.

Several recent works combine classification losses and ranking losses [Son18, Zhe18]. This can be achieved by attaching a separate classification loss to each branch of the siamese or triplet architecture.

Other Models

Besides the most frequently used classification and ranking models described in the previous sections, two other categories of models are used for more specialized applications of re-id.

Recurrent Models apply Recurrent Neural Networks (RNNs) or Long-Short Term Memory networks (LSTMs), which can capture dependencies and information across time. Consequently, these models focus on the task of re-id in videos. McLaughlin et al. [McL16] use RNNs to temporally pool information from person tracklets and learn a sequence feature through a combined classification and siamese loss function. Similarly, in [Yan16a], LSTMs are used to aggregate features over tracklets. Xu et al. [Xu17] use a siamese architecture and RNNs to fuse temporal and spatial information. Temporal information in both directions is captured by Zhang et al. [Zha18a]. Liu et al. [Liu18a] apply specialized sub-networks to consecutive frames to directly extract motion information, which is then fused with spatial information through RNNs.

In addition to extraction of appearance features, recurrent models additionally have the potential to extract and use gait information for the matching process [McL16, Liu18a]. In [Var16b], Varior et al. incorporate LSTM modules into a siamese network. Rather than using temporal information, this approach focuses on sequentially processing different parts of a single image for matching persons.

Generative Adversarial Models (GANs) are a type of CNN with the ability to generate or modify images. They are becoming increasingly popular and have been used for re-id for several different applications. Zheng et al. [Zhe17b] propose the first approach to use GANs for re-id. They generate additional unlabeled person images using a GAN and assign artificial labels with equal value for each person identity. Although the visual quality of the generated samples is low, an improvement in re-id accuracy can nevertheless be achieved. In [Wei18a], Wei et al. generate images of the same person in different poses to enhance the amount of training data. Two recent works [Zho18, Den18] show the potential of style transfer GANs to re-id for cross-domain learning or unsupervised domain adaptation. Person images are projected from a labeled source domain to a target domain in which no identity labels exist. Supervised models for the target domain can then be trained on the projected images using the labels from the source domain.

2.1.1.4 Use of Attributes

Attributes have been of interest in person re-id for some time, due to their directly interpretable semantics and demonstrated complementary information to conventional image features [Lay12a, Lay12b, Lay14, Zhu15, Su15, Su17b]. Initial approaches rely on a combination of attributes with conventional image features through application of a combined distance measure [Lay12a, Zhu15], multi-task learning [Su15], learning of a latent attribute space [Su17b], or pre-training of attribute representations of fashion data, which can then be transferred to surveillance data for re-id [Shi15].

Several recent works focus specifically on combining attribute information with CNNs. Khamis et al. [Kha14] use a triplet loss architecture for re-id

in combination with an attribute loss and leverage multiple data sources. In [Mat16], fine tune CNNs for attribute recognition and employ metric learning for subsequent person re-id. Recently, Lin et al. [Lin17] used a combination of re-id and attribute classification losses to learn a joint representation for person re-id. In [Su18], a three-stage process is proposed where attributes are first learned on separate data, then fine-tuning of attributes for re-id is performed on a re-id dataset, and finally a combined fine-tuning stage of attribute classification is carried out on the combined data. Wang et al. [Wan18b] learn a joint attribute and identity feature space, which is transferable to new target domains.

2.1.1.5 Use of Pose and View Information

Variation in a person's body pose can significantly alter visual appearance and change the location of important cues for re-id inside the image. Thus, including body pose or at least view angle information in a re-id approach can often significantly improve the resulting feature representations. Several approaches use pose information to design or learn improved features for re-id. The popular SDALF descriptor by Farenza et al. [Far10] uses two axes dependent on the body's pose to derive a feature description with a degree of pose invariance. Cho et al. [Cho16] define four view angles (front, left, right, back) and learn corresponding matching weights to emphasize matching of same-view person images. A more fine-grained pose representation based on Pictorial Structures was first used in [Che11] to focus on matching between individual body parts. More recently, the success of deep learning architectures in the context of re-id has lead to several works that include pose information into a CNN-based matching. In [Zhe17a], Zheng et al. propose to use a CNN-based external pose estimator to normalize person images based on their pose. The original and normalized images are then used to train a single deep re-id embedding. A similar approach is described by Su et al. in [Su17a]. Here, a sub-network first estimates a pose map which is then used to crop the localized body parts. A local and a global person representation are then learned and fused. Pose variation has also been addressed by explicitly

detecting body parts through detection frameworks [Zha17b] or through keypoint detection [Wei17]. Li et al. [Li17b] use Spatial Transformer Networks (STN) to learn and localize body parts in a CNN, which are combined with a global feature through concatenation. Less explicit approaches rely on visual attention maps [Rah17], or body part specific attention modeling [Zha17c]. A recent work by Zheng et al. [Zhe17b] uses GANs to generate images of the same person in new poses. The approach allows to create a dataset with uniformly distributed poses. The authors show that a lack of bias towards specific poses can lead to a notable improvement in re-id accuracy.

2.1.1.6 Re-Ranking

Re-ranking approaches are based on the initial ranking result achieved by any re-id method and aim to improve on this initial ranking. For this purpose, further information is gathered for each element in the ranked list and new distance values are computed, which often lead to an improved ranking. Re-ranking methods work fully automatically and do not require any user interaction or feedback. Re-ranking can be applied to the entire rank list or just to first ranks, depending on available resources and time.

A common strategy for re-ranking is to rely on the top-k nearest neighbors (k-NN) of each entry in the initial ranked list as additional information. For re-ranking, pairs of images are then compared according to the similarity between their neighborhood rank lists. The type of information or features based on which such nearest neighbors are determined can vary widely between approaches. Shen et al. [She12] proposed one of the first approaches to use k-nearest neighbors and produce new rank lists based on these. Other works propose to jointly learn direct image content and context information, i.e. neighborhood relationships, to remove candidates in the top neighbors [Gar15] or revise the initial ranking with a new similarity obtained from fusion of content and contextual similarity [Len15]. Relative information of common nearest neighbors is first used for re-ranking by Li et al. in [Li12a]. Ye et al. [Ye15] determine common nearest neighbors based on global and

local features as new queries and revise the initial ranking list by aggregating these into new ranking lists. Finally, in [Ye16], the use of similarity and dissimilarity cues from neighbors of different baseline methods is proposed.

In contrast to common neighbors, several recent works [Jeg07, Qin11, Zho17a] use reciprocal neighbors, i.e. common neighbors that reciprocate with each other in a k-neighborhood. Most recent state-of-the-art re-ranking methods are based on computing these neighborhood list comparisons using a generalized Jaccard distance. To overcome the associated complexity of computing intersection and unions of underlying variable-length lists, Sparse Contextual Activation (SCA) [Bai16] encodes the neighborhood set into a sparse vector and then computes the distance. A popular recent method by Zhong et al. [Zho17a] uses k-reciprocal lists and computes the Jaccard distance using an SCA encoding. This distance is then fused with the distance from the original ranking to obtain the final re-ranking. A recent approach by Sarfraz et al. [Sar18b] relies on an expanded neighbor set and defines a new distance for image pairs based on such sets. This expanded cross-neighborhood (ECN) re-ranking approach does not require expensive re-computation of rank-lists for each pair of images.

In the context of this thesis, the recent state-of-the-art re-ranking methods k-reciprocal neighbors [Zho17a] and ECN [Sar18b] are used to further improve re-id accuracy based on the initial results achieved by the proposed models.

In summary, most recent re-id methods rely on CNN-based models. Classification models and ranking models represent the two largest groups and are generally equally powerful, although ranking models often require a little more care with sample selection in the training process. Recent developments include combination of loss functions, inclusion of pose or body part localization, and the use of attention mechanisms. Attribute information is occasionally used and adds a more explicitly semantic aspect to re-id. Re-ranking methods are frequently used in addition, which can improve ranking results independently of the actual re-id approach.

2.1.2 Datasets

The number of available datasets for re-id is large. With the onset of deep learning and an increased research interest in the subject, the number and size of datasets has increased significantly in recent years. This section will review only the most well known or largest datasets. An overview is given in Table 2.1.

Early re-id datasets are usually small, containing few person IDs and often only two images of each person, recorded by two different cameras. An example for this is the classical and most well studied re-id dataset, the VIPeR dataset. It contains 632 identities, and two images for each identity. The evaluation protocol requires to define 10 random train-test splits with 316 identities in each set. Performance is then averaged across all splits. While VIPeR consists of outdoor data, other classical datasets cover a range of scenarios, including subway stations (GRID), airports (iLIDS), university campuses (CUHK01), or shopping malls (CAVIAR).

Table 2.1 shows significant growth of datasets in recent years, in terms of person IDs, number of images, and number of cameras or available images per person. This is a promising trend, which has enabled widespread use of deep learning methods in the field. Evaluation protocols have shifted to a predefined set of re-id queries and a fixed gallery in which further instances of the query persons have to be found. The conventional n-fold random split is not practical for evaluation of CNN models, as re-training n times is time consuming and the increased size of recent datasets allows for large test sets, which provide a sufficiently strong basis for stable empirical results. Another recent development is the inclusion of bounding boxes, which are automatically generated by person detectors. Early datasets often contain hand-drawn bounding boxes, which are much more accurate than can be expected in a practical system. The inclusion of automatically generated bounding boxes, often by application of the popular DPM detector, thus helps in training more robust and practically applicable models.

The two datasets, which are currently most frequently used in research, are the Market-1501 and DukeMTMC-reID datasets. The Market-1501 dataset

contains images from 6 cameras placed on a university campus. In total, 1,501 persons are recorded and split equally between the training and test split. The boxes are generated with the DPM detector and annotated manually for person IDs. For testing, the dataset contains additional distractor images in the gallery, which fit to none of the test queries. The distractors include actual person detections, as well as body-part detections, and even false positive detections. The DukeMTMC-reID dataset has similar properties as the Market-1501 dataset. Example images from both datasets, including distractor images, are depicted in Figure 2.4.

Figure 2.4: Example images from the Market-1501 and DukeMTMC-reID datasets. Rows show images of the same person. The final row shows distractor images from each dataset.

Another recent development is the emergence of datasets that provide more realistic, special scenarios for re-id. The Market-1501 dataset provides an additional set of 500,000 distractor images, which can be used to better evaluate the robustness of re-id methods for very large gallery sizes. Several datasets offer tracklets, which allow for the evaluation of re-id methods in videos, instead of single images, see Figure 2.5. The most prominent candidates here are MARS, which is based on the same data as the Market-1501 dataset, and DukeMTMC-VideoReID, which is based on the same data as the DukeMTMC-ReId dataset. The PRW and CUHK-SYSU datasets offer full camera images, instead of pre-cropped person patches. This enables the evaluation of combined approaches for person detection and re-id. In these datasets, query images are provided pre-cropped but gallery images are full camera images. For re-id, persons first have to be detected in the gallery images and these detections are then matched to the query images.

In this thesis, the popular Market-1501 and DukeMTMC-ReId dataset are used for evaluation of the proposed methods. In order to judge accuracies under conditions that more closely resemble practical application, the Market-500K, MARS, PRW, and CUHK-SYSU datasets are further used.

Figure 2.5: Example tracklets from the video-based MARS dataset.

2.1.3 Metrics

The accuracy of a re-id approach is evaluated by the quality of the resulting ranked lists. Specifically the rank positions of correct matches are assessed. The most well known metric for this is the Cumulative Matching Characteristic (CMC). Across a number of queries, the CMC reports the average probability of having encountered a correct match at each rank.

$$CMC(k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\sum_{j=1}^{k} r_{ij} \geq 1} \tag{2.2}$$

where N is the number of queries used for evaluation and r_{ij} is equal to 1, if the rank list for query i contains a correct match at rank j, and otherwise 0. The CMC can be visualized as a curve over all ranks k starting from the average frequency of encountering a correct match at rank 1 to a maximum value of 1, which is achieved at the final rank K or earlier. Due to the often small differences between results of different methods and the redundant information in exhaustively plotting CMC for all ranks, in practice rank accuracies are often reported in tables. A frequent choice are rank accuracies at ranks 1, 5, 10, and 20. A rank-k accuracy corresponds to the average frequency of queries, which contain at least one correct match between ranks 1 and k, and is thus equivalent to $CMC(k)$.

While rank accuracies or CMC give a reasonable and easily interpretable impression of the accuracy of a re-id approach, they were originally motivated by use in ranking tasks, where only one correct result was present in the ranking. Thus, for rankings with several correct results, they do not take into account the distribution of additional correct matches after the first one. Furthermore, rank accuracies and CMC do not represent a single, unified measure by which to compare approaches. The mean average precision (mAP) is thus often used in addition to rank accuracies. It is defined as

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i, \quad AP_i = \frac{1}{\sum_{j=1}^{K} r_{ij}} \sum_{j=1}^{K} r_{ij} \frac{\sum_{l=1}^{j} r_{il}}{j} \tag{2.3}$$

where again r_{ij} is 1, if the rank list of query i contains a correct match at rank j, and otherwise 0. K represents the length of a ranked list, i.e. the number of samples in the gallery. The average precision (AP) takes multiple correct matches into account by computing a precision value at each position in a ranked list, where a correct match is encountered, and normalizing the sum of these precision values by the overall number of correct matches within the given ranked list. Thus, if all correct matches are consecutively at the first positions of the ranked list, i.e. an ideal result is achieved, the precision at each position is 1 and the AP is also 1. If this is achieved for all queries considered during evaluation, the mAP has a maximum possible value of 1.

Table 2.1: An overview of classical and recent re-id datasets. †: For video-based datasets, the number of images refers to tracklets instead of single bounding boxes. ∗: In datasets with full camera images, the hand-drawn annotations are used for evaluation only. Methods have to generate their own detections, which are matched against annotations. The number of distractors in the full images is not known. The CUHK-SYSU dataset is recorded by mobile cameras, a fixed number of cameras is thus not given.

Dataset	Year	IDs	Images	Distractors	Cams	Label	Type
VIPeR [Gra08]	2007	632	1,264	0	2	hand	crop,single
GRID [Loy09]	2009	1,025	1,275	775	2	hand	crop,single
iLIDS [Zhe09]	2009	119	476	0	2	hand	crop,single
CAVIAR [Che11]	2011	72	1,220	0	2	hand	crop,single
PRID [Hir11]	2011	200	1,134	0	2	hand	crop,single
CUHK01 [Li12b]	2012	971	3,884	0	2	hand	crop,single
CUHK02 [Li13]	2013	1,816	7,264	0	10	hand	crop,single
CUHK03 [Li14]	2014	1,360	13,164	0	10	DPM/hand	crop,single
Market1501 [Zhe15b]	2015	1,501	32,217	2,793 (+500k)	6	DPM [Fel10]	crop,single
DukeMTMC-ReID [Ris16, Zhe17b]	2017	1,404	36,411	408	8	hand	crop,single
MSMT [Wei18b]	2018	4,101	126,441	0	15	Faster R-CNN	crop,single
MARS [Zhe16a]	2016	1,261	20,478†	3,248†	6	DPM+GMMCP	crop, tracks
DukeMTMC-VideoReID [Wu18]	2018	1,404	4,832†	408†	8	hand	crop, tracks
PRW [Zhe16b]	2016	932	11,816*	unknown	6	hand*/DPM	full-img, single
CUHK-SYSU [Xia16b]	2016	8,432	18,184*	unknown	variable	hand*	full-img, single

2.2 Attribute Recognition

Attribute recognition is a less popular computer vision task compared to re-id. Nevertheless, in recent years a growing number of approaches and datasets focus specifically on the topic. Since attributes are often only discernible by small visual details, the challenges faced by attribute recognition approaches are similar to those of re-id methods. An added challenge is posed by the strong imbalances in the data. While re-id datasets often contain a similar number of images for each person ID, attribute recognition datasets often have very strong imbalances between the number of positive samples for the different attribute classes. Furthermore, each attribute generally has a much larger number of images where the attribute is absent, i.e. negative samples, than images where the attribute is present, i.e. positive samples. This section first gives an overview of the related literature on the subject of attribute recognition and then describes existing datasets, as well as metrics used for evaluation of approaches.

2.2.1 Existing Methods

Attribute classification is a multi-label classification task. A straightforward way to address this is by relying on the extensive single-label classification literature and training a separate classifier for each attribute. Several early works follow this direction. Sharma et al. [Sha11] apply this approach by using spatial histogram features in conjunction with a maximum margin optimization to learn each of the attribute classes. Similarly, Layne et al. [Lay12a] and Deng et al. [Den15] use Support Vector Machines (SVMs) and a set of color and texture features to classify each attribute. Due to the independent classification of attributes, such approaches cannot directly leverage semantic relations between attributes. Thus, in [Bou11] an additional layer of SVM classifiers is employed to take these relationships into account. Following the same motivation, Chen et al. [Che12] refine initial attribute predictions through a Markov Random Field (MRF), which models attribute relationships. However, classification of attributes by individual classifiers does not scale

well for larger amounts of attributes. More recent approaches thus often rely on a single model to recognize all attributes. The majority of these are based on deep learning methods and can be grouped into several categories.

Global methods focus on directly processing the entire person image. Initial models rely on features extracted by CNNs, such as AlexNet, and classify attributes by using one loss function for each attribute [Sud15]. While this approach allows for better scalability, it does not directly solve the issue of attribute relationships. To address this, the MTCNN model [Abd15] introduces a shared layer before the separate loss functions, which helps to propagate information between the different attribute branches. A joint loss function for an entire set of attributes was first applied in [Li15]. The sigmoid cross-entropy loss is used and specifically modified to address the strong imbalances between different attributes. This joint loss is shown to result in better accuracies than individual loss functions. A similarly adapted loss function is employed in [Yu16]. Additionally, the authors use a multi-scale feature representation by relying on side-branches from different depths in the network. The accuracies achieved by global methods suffer from the fact that attributes are often located in small image regions, which may not be well represented in the internal feature space of a global CNN.

Part-based methods address this issue by including local parts into the classification model. This can be achieved by relying on external body part detectors and learning an attribute classifier for each body part [Zha14] or by diving the image into a fixed grid of patches [Zhu15]. In contrast to this, the AAWP model proposed by Gkioxari et al. [Gki15] trains both a part detector and an attribute classifier model. Similarly, in [Dib16] discriminative image patches are mined based on clustering and attribute classifiers are trained on the resulting patches. While these models are trained for patch localization and attribute classification on the same data, the two steps are not trained jointly. The first jointly trained end-to-end model for part-based attribute recognition is proposed in [Yan16b] and consists of a part-bounding-box generator and a set of separate softmax classification losses for the attributes. A very similar end-to-end approach is proposed in [Li18a], where parts are detected in

part of the network and attribute classification is carried out on a concatenated feature representation of all patches using a multi-class cross entropy loss. Besides local parts, scene context information is additionally included in model proposed by Li et al. [Li16b]. Local part information, global information, and scene information are fused by addition in a single network and used for attribute recognition through a single multi-class loss function. While all of these methods rely on local parts for attribute recognition, most of them contain a global network branch or a part that represents the whole image. Thus, the combination of local and global information is an important aspect of robust attribute recognition methods.

Attention-based methods follow a similar motivation as part-based methods. Attention mechanisms provide a convenient option to force a network to focus on certain important local areas. In contrast to part-based approaches, an attention mechanism can be more flexible in that it has more freedom in learning which areas are relevant for a given task. Furthermore, attention mechanisms do not require the hard and potentially error-prone decision of cropping image regions, which is required in patch-based approaches. Only a few approaches thus far employ attention mechanisms for attribute recognition. The HydraPlus-Net [Liu17b] applies an attention mechanism at three different levels in the network, which represent different levels of semantic information. A similar approach is proposed in [Sar18a]. Here, an attention mechanism is applied at two different locations within the network. In contrast to the HydraPlus-Net, attention is applied directly on the feature map based on which it was computed. In [Guo17] the attention mechanism is separated into a sub-network. Attention is computed based on the final layer of the main attribute branch. Due to the low spatial information in that final layer, a refinement of the attention map is applied and a specialized loss function measures the suitability of the resulting attention map. In all of these cases, self-attention is applied, i.e. the attention mask is computed directly from layers of attribute recognition network. Similar to part-based methods, all attention approaches combine the attention mechanism with a global feature branch.

Recurrent methods employ sequential CNN models, such as RNNs or LSTMs. The use of such models often aims at better capturing relationships, such as co-occurrences, between different attributes by either applying a recurrent network for all attributes globally [Wan16] or an individual recurrent element for different body regions [Wan17, Zha18b]. However, the latter two approaches rely on predetermined separations of body regions, which, similar to the previously described patch-based methods, can have a negative impact on performance. Thus, in [Liu18b] a joint approach is proposed, which combines an attention model with recurrent network elements. A main drawback of recurrent models for attribute recognition is their runtime, due to the staged attribute prediction.

A number of further methods exist, which consist of more unique approaches and do not easily fit into the larger categories. This includes the PatchIt approach [Sud16], which improves pre-training of networks by cropping person images into patches and pre-training the network to determine the original location of patches within the person image. In [Lu17b] a method is proposed which gradually splits network elements during multi-task attribute recognition, so that attributes with high synergies remain in the same layers and attributes without synergies get split into different network branches. The methods described in [Don17, Sar17a] employ curriculum learning, which focuses on learning easy tasks first and gradually add to the difficulty. Generative Adversarial Networks (GANs) have recently been applied for attribute recognition in order to increase image resolution and de-occlude body parts [Fab17].

In summary, attribute recognition approaches have moved from purely global approaches which address each attribute separately to methods, which combine global and local information and address relationships between attributes by either sharing elements of the model between all attributes or jointly predicting all attributes. Attention mechanisms are a promising direction, since they avoid hard decisions of cropping out image regions, which may lead to missed visual details.

2.2.2 Datasets

A growing number of computer vision datasets focus specifically on person or pedestrian attribute recognition. An overview is given in Table 2.2 The number of annotated person images, as well as the number of annotated attributes, has increased significantly over time. While initial datasets contained only a few thousand images and tens of attributes, more recent datasets approach 100,000 images with more than 60 annotated attributes.

Most attributes are annotated as binary values, i.e. present or not present. Examples for such attributes are male, long-hair, hat, sunglasses, pullover, jeans. Some dataset additionally offer multi-valued (multi-class) attributes. These most frequently include colors, textures, or in some cases the orientation of the person towards the camera (view angle). Attributes from the existing datasets can be grouped into several categories:

- **Soft-Biometrics** refer to weak biometric cues that describe a person's identity. These are attributes that are directly associated with the person and not easily changed. Examples include gender, hair color, hair style, skin color.
- **Clothing Attributes** are by far the largest category of attributes and describe a person's clothing types, colors, textures, or accessories. Examples include hat, sneakers, jeans, pullover, upper-body-color, lower-body-color, hand-bag, suitcase.
- **Activity Attributes** Some datasets, such as RAP, additionally provide attributes relating to a person's acitivity. Examples include talking, carrying, or telephoning.
- **Situational Attributes** refer to additional properties specific to the current recording conditions. Primarily, this includes the view angle of the person or annotation of occluded areas.

Besides their semantic differences, these categories also reflect the temporal persistence of the different types of attributes, which is important for their application in re-id . Soft biometrics can be expected to remain the same during the timeframe relevant for most re-id applications. Clothing attributes are

more easily altered but often remain stable. Activity attributes may change frequently and situational attributes are almost certain to change between different recordings.

Several attribute recognition datasets are at least partly created from re-id datasets. This includes the PETA, Market-1501, and Duke-MTMC datasets. This constellation allows for training of joint attribute and re-id models on the same dataset. However, such models learn to expect an unusually high accuracy in attribute predictions and thus do not transfer well to other datasets. Furthermore, in practical applications, attribute annotations are not readily available and very costly to annotate. Thus, an attribute recognition model will have to be trained on separate data and when used for re-id, the re-id approach will require a mechanism to handle potential errors in the attribute predictions.

An important issue with binary annotated attributes is annotation uncertainty. As attributes can often be difficult to make out in low resolution surveillance imagery, annotators are often unsure, if an attribute is actually present. In several cases wrongly cropped person images, viewpoints, or occlusions can lead to attributes not actually being visible in the image. Most datasets nevertheless require a binary choice, which leads to noisy annotations. Exceptions to this are the PARSE27K, APiS, BAP, and CAD datasets, which contain a third possible attribute value to mark uncertainty or non-visibility. In cases where person IDs are available in the data, e.g. Market-1501 or Duke-MTMC, attributes are often annotated at ID-level. It may, however, be the case, that different recordings of the same person show different attributes, as not all attributes are persistent over time. Thus, annotation quality of attribute datasets is often lower than that of re-id datasets.

In the context of this thesis, the PETA, RAP, WIDER, and PA-100k datasets are chosen for training and evaluation of the proposed methods. These choices are motivated primarily by the large number of available attributes (PETA, RAP) or the amount of available images (PA-100K). The majority of datasets focuses on surveillance data and is thus biased towards the typical upright body pose of pedestrians. The WIDER dataset is additionally chosen because it contains a wider range of body poses and thus allows for a better impression

of the proposed method's accuracy under this additional challenge. Several example images of these datasets are depicted in Figure 2.6.

Figure 2.6: Example images for some of the attributes of the PETA, RAP, and WIDER datasets.

2.2.3 Metrics

One of the most widely used metrics to evaluate the accuracy of attribute recognition methods is the mean accuracy (mA). Due to the imbalances in positive and negative values for most attributes, these are considered separately:

$$mA = \frac{1}{2C} \sum_{i=1}^{C} \left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i} \right) \tag{2.4}$$

where C corresponds to the number of attributes and TP_i and TN_i are the number of correctly predicted positive and negative examples, respectively. P_i and N_i refer to the total number of positive and negative examples in the

Table 2.2: Overview of public person attribute datasets. The datasets contain binary valued attributes (present, not present) or multi-class attributes, such as colors. Most datasets consist of surveillance data and some contain additional person ID annotations.

Dataset	Year	Images	Attributes	Domain	Re-Id
HAT [Sha11]	2011	9,344	27 binary	Web	
BAP [Bou11]	2011	8,035	9 binary	Photo	
CAD [Che12]	2012	1,856	23 binary, 3 multi-class	Web	
APiS [Zhu13]	2013	3,661	11 binary, 2 multi-class	Vehicle	
PETA [Den14]	2014	19,000	61 binary, 4 multi-class	Surveillance	✓
PARSE-27K [Sud15]	2015	27,000	8 binary, 2 multi-class	Surveillance	
CRP [Hal15]	2015	27,454	1 binary, 13 multi-class	Surveillance	
RAP [Li16a]	2016	41,585	69 binary, 3 multi-class	Surveillance	
WIDER [Li16b]	2016	13,789	14 binary	Web	
PA-100K [Liu17b]	2017	100,000	26 binary	Surveillance	
Market-1501 [Lin17]	2017	32,668	26 binary, 1 multi-class	Surveillance	✓
DukeMTMC [Lin17]	2017	34,183	23 binary	Surveillance	✓

data. Considering accuracy of negative and positive values separately penalizes models, which are biased to predict only the more frequently occurring of the two values. Due to the strong imbalance between the values, a constant negative prediction would otherwise result in a high accuracy. In addition to mA, a corresponding version of mAP can be applied for attribute retrieval as well:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i, \quad AP_i = \frac{TP_i}{PP_i} \tag{2.5}$$

where PP_i is the number of positive predictions for attribute i.

However, these metrics regard attributes independently of each other and do not take correlations or relationships between attributes into account. A method with high mean accuracy may thus still generate attribute descriptions of people that contain strong inconsistencies. To address this issue, Li et al. [Li16a] introduced several new metrics, which aim at better representing the semantic consistence of attributes for an average sample image. Since these new metrics focus on correctness of the attribute description for entire images, they are termed *example-based* metrics, while measures treating attributes independently are called *label-based*. The example-based metrics adapted in [Li16a] are well known from other computer vision tasks, such as object detection, and include accuracy, precision, recall and F1 score:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i^+ \cap f_{att}^+(x_i)|}{|Y_i^+ \cup f_{att}^+(x_i)|} \tag{2.6}$$

$$Prec = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i^+ \cap f_{att}^+(x_i)|}{|f_{att}^+(x_i)|} \tag{2.7}$$

$$Rec = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i^+ \cap f_{att}^+(x_i)|}{|Y_i^+|} \tag{2.8}$$

$$F1 = \frac{2 * Prec * Rec}{Prec + Rec} \tag{2.9}$$

where N represents the number of images, Y_i^+ the positive labels of the i-th image, and $f_{att}^+(x_i)$ the predicted positive attributes for this image. $|\cdot|$ denotes the set cardinality.

2.3 Deep Learning

Neural networks have been used in computer vision and related search areas for a long time. They have progressed through several evolutionary stages over the years and their most recent form is often termed Deep Learning [Sch15a], due to the large number of network layers that have become possible with the advancement of the required hardware. The most popular variant of neural network currently applied to many computer vision task is the Convolutional Neural Network (CNN). The widespread use of CNNs started with the AlexNet model [Kri12], which, in 2012, achieved significantly improved accuracy in the ImageNet classification challenge [Den09]. The challenge requires the classification of images into one of 1000 classes of diverse objects. The success of AlexNet in this task was an initial indicator for the ability of CNNs to capture a large and diverse number of image contents and thus address a wide variety of computer vision tasks.

2.3.1 The Multi Layer Perceptron

The most common basic unit in a neural network is the perceptron [Ros58]. A perceptron accepts an arbitrary number of n input values, generally provided as an n-dimensional vector $\mathbf{x} \in \mathbb{R}^n$. The scalar output y of the perceptron is defined as a weighted sum, passed through an activation function:

$$y = \Phi(\mathbf{w}^T\mathbf{x} + b) \tag{2.10}$$

with weight vector $\mathbf{w} \in \mathbb{R}^n$, optional bias term b, and activation function Φ. The set of weights $\mathbf{w}$ and the bias value b are the free parameters learned during training neural networks. The purpose of the activation function Φ is to introduce a non-linearity, as even a combination of many perceptrons would otherwise only be capable of learning purely linear decision functions. Typical choices for the non-linearity Φ are the sigmoid function, the hyperbolic tangent, or, most frequently, the Rectified Linear Unit (ReLU), $\Phi(x) = max(0,x)$ [Kri12].

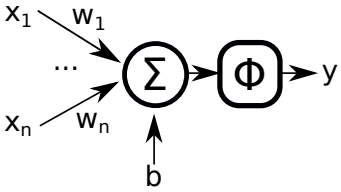

Figure 2.7: A single perceptron with input vector $\mathbf{x}$, parameters $\mathbf{w}$,b, and output y.

While a single perceptron is limited in its ability to approximate a decision function, the combination of many such units as sequential layers, called a Multi Layer Perceptron (MPL), can accurately approximate very complex functions. Within an MLP sets of perceptrons form layers, with the inputs of each perceptron in layer i connected to the outputs of all perceptrons in the previous layer $i-1$. Thus, these layers are often referred to as *fully connected* layers. The output of a fully connected layer i is defined as:

$$\mathbf{h}_i = \Phi(\mathbf{W}_i \mathbf{h}_{i-1} + \mathbf{b}_i) \tag{2.11}$$

where $\mathbf{h}_0$ would be the input to the network, e.g. an image or feature vector, and $\mathbf{h}_N$ for final layer N would be the output of the MLP. Note that the number of trainable weights in each layer corresponds to the product of outputs from the previous layer and the number of elements in the given layer. Fully connected layers can thus significantly increase the number of parameters of a neural network.

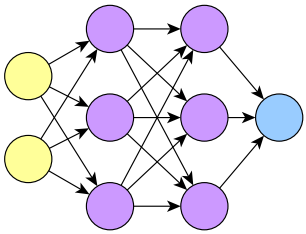

Figure 2.8: A multi layer perceptron (MLP) with two input neurons, two intermediate layers with three neurons each, and a single output neuron.

2.3.2 Convolutional Neural Networks

Convolutional network layers aim to reduce the number of parameters resulting from high-dimensional inputs, such as images. This is achieved by setting two important restrictions on the numbers and values of weights in the convolutional layer (see Figure 2.9). First, a neuron in the layer i is only connected to a local neighborhood of neurons in the previous layer. The size of this neighborhood is a design parameter of the convolutional layer. Furthermore, the weights of all neurons in the convolutional layer is forced to be the same. These two restrictions lead to a notably reduced number of weights in convolutional layers. Essentially, a convolutional layer describes a filter kernel with a size corresponding to the local neighborhood, which is applied at each location of the output of the previous layer. The local neighborhood is often referred to as the *receptive field*. Convolutional layers are well suited to process 2-dimensional representations, such as images, since they maintain the spatial structure. Often, a single convolutional operation does not capture sufficient information from the previous layer. In order to address this, a convolutional layer can consist of multiple *channels*. Each channel is free to learn a different set of weights, and thus a different convolutional operation. Due to its spatial nature, the output of a convolutional layer can be referred to as a *feature map* where the map spans across the spatial dimensions and at each location contains a feature vector with a dimension equal to the number of channels in the layer.

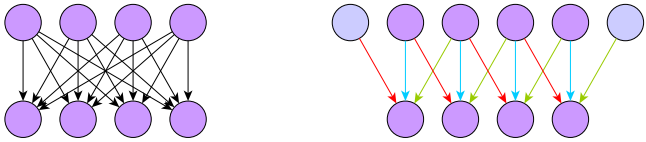

Figure 2.9: Comparison of connections, i.e. weights, of a fully connected layer (left) and a convolutional layer with a receptive field of size 3 (right). The number of weights is significantly reduced in the convolutional layer. Edge colors of the convolutional layer indicate matching weights. In practice, padding elements (light blue) corresponding to half the size of the receptive field are introduced to prevent a decrease in the output dimension of convolutional layers.

CNNs are networks comprised of one or more convolutional layers. In such networks, convolutional layers are often followed by non-linearity layers, such as ReLU, to better approximate complex objective functions. Furthermore, pooling operations, which aggregate outputs within patches across spatial dimensions are used to decrease the number of overall parameters in networks with several layers. Typically, a pooling layer is parameterized to combine a 2×2 local neighborhood into a single output value by either max- or average-pooling.

2.3.3 Loss Functions and Optimization

For training a CNN model, an objective function, or loss function, is required. The loss function compares the predicted values of the network to the ground-truth and computes an error measure. The aim of the training process is to minimize this error by use of the gradient of the loss function, which is back-propagated through all layers of the network and used to update the weights in a corresponding direction. A typical optimization method for this is stochastic gradient descent (SGD). In SGD, a small, randomly chosen, part of the training data, called a *batch*, is used to calculate an approximation of the real gradient of the objective function with respect to the current network parameters. In each iteration of the training process, a different batch is chosen and a small modification is applied to the parameters of the network, depending on the value of the approximated gradient. The *learning rate* of

the training process specifies how large this step is in relation to the gradient. In this way, the resulting values of all parameters in the network ultimately depend on the gradient of the loss function. The choice of a proper loss function can thus have a significant impact on the outcome of the training and the properties of the learned representations in the network.

Within the context of person re-identification two types of loss functions are frequently employed.

- **Classification Loss:** Classification loss functions are the established choice for many computer vision problems. In this case, the final layer of the network is designed to have as many output neurons as there are classes to be recognized. In case of the ImageNet challenge this would correspond to a 1000-dimensional final layer. For re-id the number of different persons in the training data is chosen. The standard choice for a classification loss function is the softmax cross entropy loss [Dud12].
- **Ranking Loss:** Ranking losses require more than one image as input during training of the network. Ranking losses optimize distances between feature representation of object class images. This is generally achieved my forcing the feature distances between images of the same class to be small and distances between images of different classes to be large. A typical representative of this class is the triplet loss [Wei09].

2.3.4 CNN Training Practices

The training of CNNs is a complex process which requires many design choices in order to achieve a stable convergence and high accuracies in the targeted computer vision task.

A proper initialization of network weights at the beginning of the training process is often very important. This is particularly the case, when only limited amounts of training data are available and the used network architecture contains many parameters. The weights of models trained for the ImageNet

challenge or on related very large datasets are a popular choice for initialization. The diverse set of classes in the ImageNet challenge results in models which can be adapted to a large variety of tasks. Random initialization of weights and training from scratch offers more flexibility in the design of new network architectures but often leads to less optimal results, due to convergence to local minima or unstable training caused by vanishing or exploding gradients.

Besides stochastic gradient descent with a fixed learning rate, several other optimization methods exist, which aim at providing an adaptive learning rate. A simple and frequently used addition is *momentum*, where a fraction of the previous update of a network parameter is added in the next update step. Thus, in case of consistent gradient slope over several iterations, the size of the steps taken in the direction of the gradient become increasingly larger or, in other words, gain momentum. Besides SGD with momentum, several other methods for dynamically adapting the learning rate have been proposed, including AdaDelta [Zei12], RMSprop, and Adam [Kin14]. Particularly the latter, Adam, often leads to good results in practice and avoids time consuming training of several networks with vanilla SGD and various learning rate schedules.

Due to their large number of parameters, overfitting, i.e. an over-adaptation to the observed training samples, is a common problem with CNNs. To address this, regularization methods can be employed. A frequently used method to prevent overfitting is *weight decay*. Weight decay introduces a gradual decay to each weight by reducing it by a small fraction of its current value. This prevents overly large weight values, which otherwise dominate the output of the network. Another popular option to regularize neural networks is dropout [Sri14]. Here, neurons of certain layers are deactivated during training with a specified probability. This prevents forming of critical paths through the network and induces the network to learn redundant representations for the given task. Dropout is best suited for application in network regions where most of the parameters are located, i.e. usually in fully connected layers near the end of the network.

With increasing depth of a network, the gradient signal can become less stable, leading to vanishingly small or exploding gradient values. To prevent this and achieve a more stable training of deeper and more expressive networks, *batch normalization* has been proposed [Iof15]. Layer outputs are normalized through the learned parameters α and β:

$$\mathbf{h}_i = \alpha_i \mathbf{h}^i + \beta_i. \tag{2.12}$$

These parameters are learned in such a way that the resulting output feature map has zero mean and unit variance. When initialized as $\alpha = 1$ and $\beta = 0$ the above equation corresponds to the identity. For application during test time, the parameters computed across the entire training population are used.

Data augmentation is a popular method to further aid in the training of CNNs. With the onset of deep learning many computer vision subject areas suffered from a lack of large datasets to successfully train CNNs. Augmenting training data by applying small transformations to the input images during training can be an efficient way to avoid overfitting to small datasets. Common data augmentation transformations, which were used in the ImageNet challenge [Kri12, Sim14], include small image translations, horizontal image flipping, or introduction of color shifts. Data augmentation does not only improve generalization capabilities of the resulting models but can also be introduced to specifically develop robustness towards the applied transformations in the resulting model. For example, artificial introduction of Gaussian noise can increase the model's ability to handle noisy surveillance data.

2.3.5 CNN Architectures

Since the development of early CNN models, such as AlexNet, several landmark architectures and design choices have been proposed. Often these methods have been the winning entries in the ImageNet classification challenge [Den09] and have subsequently been fine-tuned and adapted to a number of other tasks.

- After the initial success of the AlexNet architecture [Kri12], the **VGG** networks [Sim14] were the first to rely strongly on smaller convolutional filters (see Figure 2.10 (a)). A sequence of 3×3 convolutions can result in the same receptive field as a single larger convolution while also saving parameters. For example two 3×3 convolutions have the same receptive field as a single 5×5 convolution and use fewer parameters. The VGG architectures apply many such 3×3 sequences resulting in a single-path network with a very large number of parameters. While the VGG architecture achieved better results than previous approaches, the resulting models have large numbers of parameters which make a successful training on smaller datasets for fine-grained classification challenging.
- A Residual Network [He16], shortened **ResNet**, is a neural network architecture which solves the problem of vanishing gradients by providing elements of the network with an identity shortcut, which helps to robustly backpropagate the gradient signal (see Figure 2.10 (b)). These shortcuts also help to simplify the learning task, because the residual element only has to learn an offset to its input, due to the addition operation at the end, and no longer a full feature transformation. The shortcuts thus allow for a successful and robust training of much deeper architectures than previously possible. Convolutions of size 1×1 can be applied to reduce the number of channels inside a network element which allows for training of even deeper networks with a reasonable number of parameters. Such 1×1 convolutions are often called bottleneck layers.
- Following the idea of identity shortcuts for a more robust training, Dense Networks or **DenseNets** [Hua17] connect every layer to every other layer inside a block. The element-wise addition used in ResNets is replaced by a concatenation operation which maintains the individual information of both the input and the skipped layers (see Figure 2.10 (c)). In this way, there is always a direct route for the information backwards through the network. Similar to other

state-of-the-art architectures, DenseNets use 1×1 convolutions as bottlenecks before each 3×3 convolution to reduce the number of input channels and to improve computational efficiency.

- The **inception** module was first applied in the GoogLeNet model [Sze15] and follows ideas by Lin et al. [Lin13]. Rather than relying on the traditional approach of single-path networks which stack convolutional and pooling layers in a sequential structure without branching, GoogleNet uses a more complex deep network element, termed inception module. The inception module consists of parallel paths of 1×1, 3×3, and 5×5 convolutional filters, which are combined through concatenation (as depicted in Figure 2.11 (a)). Each branch results in a different receptive field. This multi-scale view on the input of the module allows the model to recover both local features via smaller convolutions and features with more context through larger convolutions. Bottleneck layers are used to reduce the number of channels before more complex convolutions. Later approaches combined the inception module with residual connections [Sze17] or studied variations of the inception architecture by making it deeper and wider, increasing the number of inception modules and simplifying the architecture. These modifications are combined into the Inception-v4 architecture [Sze17]. Similar to previous versions, this latest inception architecture does not use convolutions larger than 3×3, and uses factorization to replace large 7×7 filters with a pair of 1×7 and 7×1 convolutional layers.

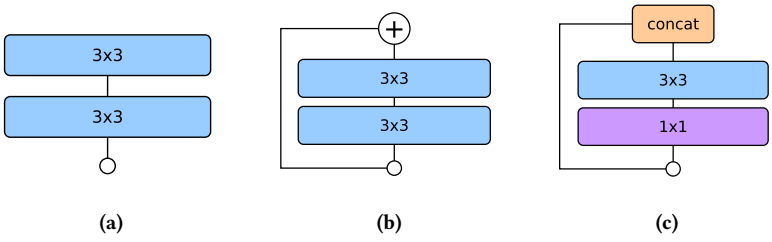

Figure 2.10: Basic building blocks of (a) the VGG architecture [Sim14], (b) the ResNet architecture [He16], and (c) the DenseNet architecture [Hua17].

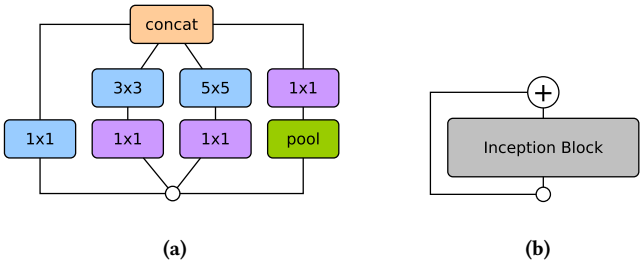

Figure 2.11: (a) An inception block as used in the GoogLeNet architecture [Sze15] and (b) a building block of the Inception-ResNet architecture [Sze17].

Chapter 3

Concept

A several stage long pipeline is required to enable person re-identification in surveillance camera footage (Figure 3.1).

Based on raw video footage, persons first have to be detected within image frames of individual cameras. Then, a tracking approach can be used to connect detections over time. The resulting person tracks are transformed into feature representations, which are finally matched to determine the visual similarity of the depicted persons. Typically, only the latter two stages are referred to as the actual re-identification (i.e. the black-dashed part in Figure 3.1). This thesis focuses primarily on those stages although the effects and potential errors, which occur in the detection and tracking stages, are addressed as well.

Two research directions are explored in this work in order to develop robust re-id models that are able to cope with errors from previous stages of the pipeline:

- The explicit *inclusion of added information* into the learning process for a re-id feature representation can help to guide and improve the learning process. In the context of this thesis, semantic attribute information, as well as person pose information have been identified

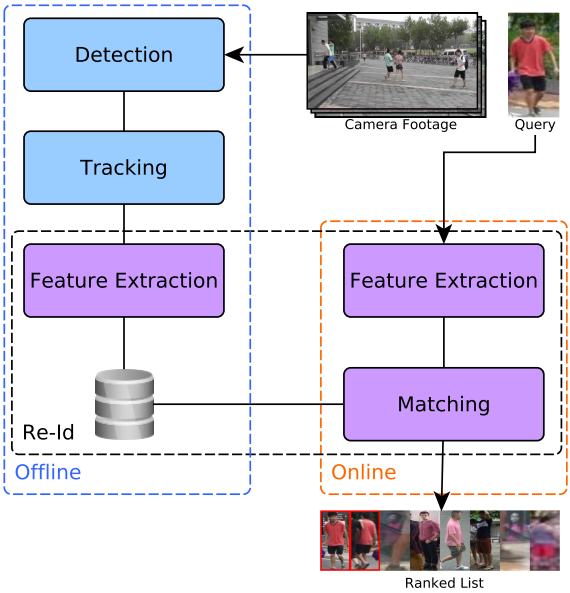

Figure 3.1: Full re-id pipeline in a practical setting. In surveillance cameras persons have to be detected and potentially tracked across time. Re-id features can then be extracted and stored. This process can happen offline, i.e. before any re-id query is made. Such a re-id query is provided in form of a person image or track. The feature representation for the query is computed and then matched against the stored database of features. The result is a list of persons ranked by similarity to the query image.

as highly useful additional information, which can increase robustness to many of the challenges discussed in Chapter 1.

- Achieving a degree of *online adaptation* of re-id methods to new scenes or cameras can greatly ease deployment of a re-id system in real world scenarios.

Figure 3.2 depicts how these research directions fit into the concept and structure of this thesis, resulting in two final models for within-dataset and cross-dataset re-id.

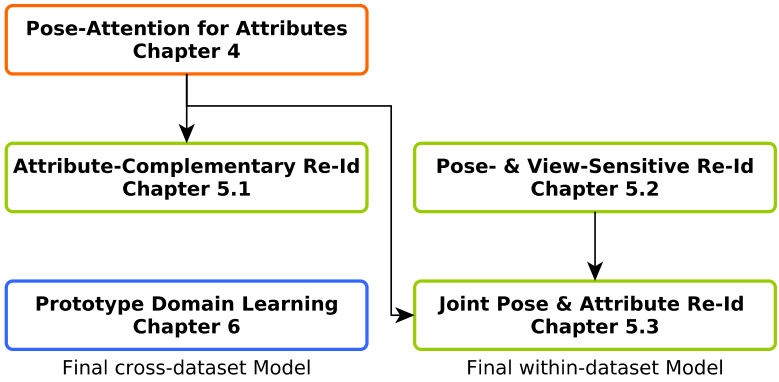

Figure 3.2: Concept and structure of the thesis. The green re-id models are developed and evaluated for a within-dataset setting, while the blue model specifically addresses the case of model deployment on data with new and unknown characteristics. The attribute model, which provides attributes or related information for re-id, is marked in red.

3.1 Semantic Attributes

Semantic attributes, often simply referred to as attributes, are a popular descriptor of a person's appearance (see Section 2.2) and several advantages make attributes a popular addition in re-id. They add a valuable semantic component to the feature representation, which is directly interpretable by humans. Thus, they can help to increase the *interpretability* of a re-id result, e.g., by providing insight into why a certain person image was ranked higher than others in the resulting ranked list of a given query. Attributes can further capture information, which may not be contained in conventional re-id features, because they often describe *small details* of persons that are only visible in localized areas of the image. Given an accurate attribute classification, the robustness of a re-id method to *illumination* and *varying sensor characteristics* can also be improved by relying on attributes.

In order to include attributes into a CNN-based re-id approach, this thesis first proposes a new model for automatic attribute recognition, which achieves state-of-the-art accuracies. The model consists of a CNN architecture which

relies on a combination of global and local information to recognize even those attributes that are only visible in very small areas of the image. The resulting attribute descriptions are then introduced into the learning process of a re-id CNN through a modified version of the triplet loss. The proposed loss function uses the attributes to guide the learning process of a re-id feature in such a way that information complementary to that already contained in the attributes is emphasized. The resulting combination of attributes and complementary features shows a clear improvement over direct re-id representation learning with CNNs, as well as conventional feature fusion methods.

3.2 Pose and View

A person's body pose can be efficiently described by a set of keypoints corresponding to the body joints in the image. Pose information provides a more precise localization of the person in an image, compared to bounding box detections. Explicit inclusion of pose information into a re-id system can help to improve robustness to *variation in body pose*. A detailed pose description can also contain information on the *camera view* relative to the person, i.e., if a person is depicted from the front or back. *Partial occlusions* by other people or scene elements can be recognized as missing keypoints in the pose representation and *small details* on limbs may be better located in the feature learning process. Explicit pose estimation in the re-id stage further helps to identify and to some extend mitigate errors caused by the previous stages of *person detection or tracking*.

Person pose estimation based on keypoints is a problem which is well addressed in existing literature. This thesis thus relies on established pose estimation models [Cao17, Ins16]. However, in addition to a keypoint based pose representation, a CNN model for estimation of a person's view relative to the camera is developed. Both types of pose information are included into a re-id CNN model to help guide the learning process. View and pose are shown to be complementary and can improve re-id accuracy significantly. The combination of pose- and attribute-based information into a final model improves accuracy even further and performs well on several realistic settings, such

as video-based re-id, large gallery sizes, and application in conjunction with person detectors.

3.3 Adaptive Re-Id

The ability to *adapt to new characteristics* of unseen scenes or changes within existing camera networks can significantly reduce the deployment effort for a real-world re-id system. The typical approach to this is unsupervised domain adaptation, where a model is adapted to new scenes by use of unannotated images, which are often readily available. However, the adaptation process may still take a lot of time and suffer from biases in the data used for the adaptation. Furthermore, frequent changes in camera positions, camera replacements, addition of new cameras, or changes in fashion trends may require renewed adaptation.

Thus, in this work an approach is proposed, which does not require any new training data and adapts itself dynamically at query level. This is achieved by selecting one of several models, which are each adapted to a specialized set of scene and person characteristics, so-called prototype domains. The adaptation requires only a matching of the query image to the model candidates and does not result in significant delays in runtime. Besides a more convenient practical use, adaptation at the level of each individual query may even improve the *scalability* to a larger number of persons in the gallery. The proposed approach is evaluated in a cross-dataset setting, which simulates deployment in new and previously unseen surroundings.

Chapter 4

Attribute Recognition

The automatic recognition of person attributes, such as gender-male, wearing-backpack, wearing-scarf, suitcase, etc., in surveillance footage is a challenging problem. As depicted in Figure 4.1, the relevant visual cues required to decide on the presence of an attribute are often only observed within small regions of the person image. The low resolution and image quality of surveillance data pose challenges. Furthermore, changes in viewpoint, a person's pose, and possible alignment errors resulting from a previous person detection stage can lead to strong variations in attribute appearance and location. This chapter describes a CNN architecture, which addresses these problems through a pose-guided attention mechanism and shows robust attribute recognition results on several public datasets.

Figure 4.1: Challenges of attribute recognition include low image quality, occlusions, high visual variation, and different location and appearance of attributes depending on view and person pose.

4.1 Pose-Attention for Attribute Recognition

Visual attention mechanisms have recently been used for several computer vision tasks, particularly those involving small image details, such as fine-grained classification, re-id, or attribute recognition. An attention mechanism is usually implemented as learning a map of weights, which are then multiplied to the main feature maps of a network. The weights range from 0 to 1 and can thus emphasize certain areas in the feature map while reducing the influence of others. Existing approaches typically employ self-attention, which relies on computing the attention map directly from the feature map it is then applied to. In contrast to this, the proposed approach relies on deriving the attention map from external information, namely from a body pose estimation result. In a practical application, such information may be readily available, if the detection stage prior to the attribute recognition model relies on a body pose detector, instead of a bounding-box based detector. The motivation for this pose-attention approach lies in the observation, that the location of most attributes does not strongly depend on the visual appearance of a person, but rather their body pose. The hypothesis is thus, that

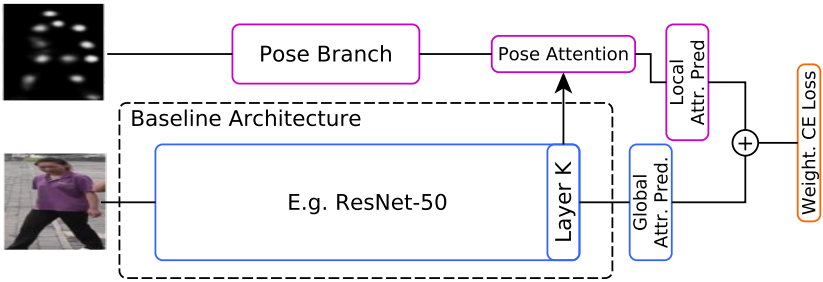

Figure 4.2: The proposed pose-attention attribute recognition architecture. As an example, the ResNet-50 CNN is used as a backbone model. The pose attention module is depicted in detail in Figure 4.3.

pose information is sufficient for attribute localization and may actually be better suited than self-attention to aid in attribute recognition. A high-level overview of the proposed attribute recognition architecture is given in Figure 4.2.

Let $\mathbf{I}$ denote an input person image with ground-truth labels $\mathbf{y} = [y_1, y_2, \dots, y_C]^T$ and $\mathbf{P}$ a suitable spatial representation of the person's pose. C corresponds to the number of attributes in the given dataset and y_i is a binary label with value 1, if the attribute is present in the image and value 0 otherwise. If the attention mechanism is to be applied to a backbone CNN after layer K, then

$$\begin{aligned} \mathbf{X}_K &= f_K(\mathbf{I}; \theta_K), \ \mathbf{X}_K \in \mathbb{R}^{n \times n \times j} \\ \mathbf{X}_{pose} &= f_{pose}(\mathbf{P}; \theta_{pose}), \ \mathbf{X}_{pose} \in \mathbb{R}^{n \times n \times l} \end{aligned} \tag{4.1}$$

where $\mathbf{X}_K$ represents the feature map output at layer K and $\mathbf{X}_{pose}$ denotes a feature derived from the pose information through a separate branch of equal spatial dimension $n \times n$. θ_K and θ_{pose} represent the parameters of the backbone architecture until layer K and the pose branch, respectively. Two separate 1×1 convolutional layers are then applied to normalize the number

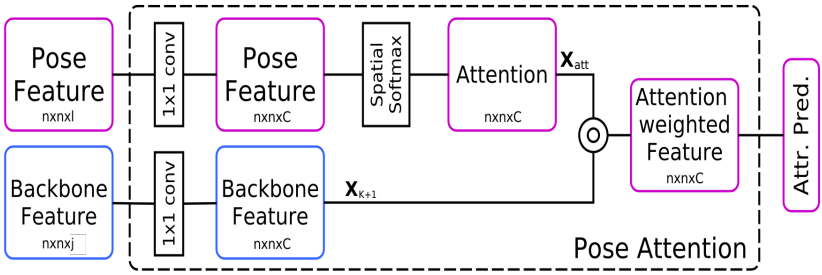

Figure 4.3: Pose-based attention block used in the attribute recognition architecture. The operator ∘ denotes element-wise multiplication.

of channels in each branch:

$$\begin{aligned} \mathbf{X}_{K+1} &= conv_{1\times1,C}(\mathbf{X}_K, \theta_{K+1}),\ \mathbf{X}_{K+1} \in \mathbb{R}^{n\times n\times C} \\ \mathbf{X}_{att} &= conv_{1\times1,C}(\mathbf{X}_{pose}; \theta_{att}),\ \mathbf{X}_{att} \in \mathbb{R}^{n\times n\times C}. \end{aligned} \tag{4.2}$$

The resulting number of channels corresponds to the number of attributes. Thus, each channel in the backbone network represents a feature relating to one of the attributes and each channel in the pose branch represents the raw attention values for that attribute feature map. The attention values must still be normalized using a spatial softmax at each channel c by

$$\mathbf{X}_{natt}(i,j,c) = \frac{\exp(\mathbf{X}_{att}(i,j,c))}{\sum_{h=1}^{n}\sum_{w=1}^{n}\exp(\mathbf{X}_{att}(h,w,c))}, \tag{4.3}$$

and are then applied through element-wise multiplication to the backbone feature maps:

$$\mathbf{X}_{K+2} = \mathbf{X}_{K+1} \circ \mathbf{X}_{natt},\ \mathbf{X}_{K+2} \in \mathbb{R}^{n\times n\times C} \tag{4.4}$$

The spatial softmax operation normalizes each channel's values to sum up to 1 and thus limits the budget of attention that can be paid across the spatial dimensions. The attention module is schematically depicted in Figure 4.3. The resulting feature maps are then connected to an attribute classification layer.

The attention-based attribute classifications are combined with a global attribute classifier.

4.2 Choice of Loss Function

Attribute recognition is a multi-label recognition task. In contrast to conventional classification tasks, more than one class label can be true for the same image. Thus conventional loss functions, such as softmax could only be applied separately for each attribute. This approach does not scale well when the number of attributes is large. Thus, a cross-entropy loss is usually applied. However, additional problems are the large imbalances for attribute labels in common datasets. In most cases, any given attribute c has far fewer images in which it is present, i.e. $y_c = 1$, than images in which it is absent, i.e. $y_c = 0$.

To handle such imbalances in the data, following [Li15], a modified weighted cross-entropy loss is applied at the final layer:

$$L_{attr} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c(y_{ic})log(\hat{y}_{ic}) + (1 - y_{ic})log(1 - \hat{y}_{ic}) \tag{4.5}$$

where $w_c = exp(-a_c)$ is the weight for c-th attribute and N the number of images. a_c is the frequency with which the c-th attribute appears in the training set. $\hat{y}_{ic}$ is the predicted probability for the c-th attribute of the image $\mathbf{I}$. Through this modification, the loss function assigns a high weights to cases when the image contains an attribute, which is very rare in the data. The network is thus more strongly penalized, if it fails to correctly classify these rare cases. Class imbalances are typically addressed through data augmentation of under-represented classes. However, this is not possible in the case of attributes, as balancing one attribute will introduce new imbalances on others. The weighting of individual attributes in the loss function is a prevent this.

4.3 Implementation Details

The pose information used to extract the pose-attention is based on the DeeperCut pose estimator [Ins16]. Rather than relying on the final pose output, the previous layer's pose probability maps are used, as they provide a more accurate picture of possible joint locations and estimation certainty. The pose information is provided to the pose branch as a 14-channel input map, one channel for each available joint location. The pose branch itself consists of a lightweight sequence of convolutional layers with ReLU activations and BatchNorm, paired with an equal number of pooling layers as in the backbone network.

To prevent overfitting and achieve a more robust attribute recognition, data augmentation is applied during training. Input images are first normalized to zero-mean and resized to 10% above the network's expected input size. For batch creation we then randomly crop the images down to the required size and apply random horizontal flipping. The pose branch's input undergoes the same transformations. All network's weights are initialized from a pre-trained ImageNet model and fine-tuned with an initial learning rate of 0.001 using the Adam solver with a batch size of at least 32, depending on the size of the network in memory. All added elements of the network, i.e. the pose branch and the classification layers, use a learning rate that is by factor 10 higher than the rest of the network.

4.4 The VeSPA Model

Prior to the proposed pose-guided attention model, a view-sensitive pedestrian attribute recognition approach (VeSPA) [Sar17b] was developed in joint and equally contributed work with Saquib Sarfraz. This model relies on global view information, i.e. frontal, side, back, and integrated it into an attribute recognition CNN. While the model was originally developed for attribute

recognition, it showed greater promise in application to re-id and is thus described in more detail in Section 5.2.1. The two models have the use of additional information in common, i.e. pose or view, but they differ significantly in the way this information is used. The VeSPA model applies view information to globally weight entire feature maps, independently of individual attributes. In contrast, the pose-guided attention model learns separate information for each attribute and applies it locally to specific spatial regions within feature maps. The latter approach better incorporates the observed trend of combining local and global information for accurate attribute recognition, see Section 2.2.1. Furthermore, while the VeSPA model learns specialized feature representations for different views, the proposed pose-guided attention model learns specialized feature representations for each attribute, which aligns more directly with the underlying task.

4.5 Evaluation

In this section, the proposed Pose-guided Attribute Attention model (PGA) will be evaluated on several public datasets. A comparison to state-of-the-art methods will be carried out and a dataset suitable for pre-training the attribute model prior to use in re-id will be identified. Finally, several ablation experiments are performed to gain insight into the impacts of backbone architecture, weighted attribute loss, and attention mechanism on the attribute recognition accuracy. Unless otherwise specified, ResNet-50 is used as the backbone architecture.

Datasets: The PGA model is evaluated on the PETA, RAP, WIDER, RAP2, and PA-100k datasets, which feature different sets of attributes.

The PETA dataset consists of 10 small, publicly available surveillance datasets from the fields of person re-id, detection, or tracking. The dataset consists of 19,000 images, each annotated with 61 binary and 4 multi-class attributes. PETA it is randomly partitioned into 9,500 for training, 1,900 for verification and 7,600 for testing. Following the established protocol, 35 attributes for

which the ratio of positive examples is higher than 1% are chosen for evaluation. Nevertheless, the data imbalance is so severe that the positive ratios of attributes vary from the lowest 1.2% of V-neck, to the highest 86.1% of attribute causal lower body clothes. Due to its 10 data sources, the PETA dataset cover a good range of visual variation and should be well suited to train models, which generalize well to new data.

The RAP dataset is collected from real indoor surveillance scenarios and 26 cameras are selected to acquire images, it contains 41,585 samples with resolution ranging from 36×92 to 344×554, Specifically, there are 33,268 images for training and the remains for testing. 72 fine-grained attributes (69 binary attribute and 3 multi-class attribute) are assigned to each image of this dataset. Following established protocol, 51 binary attributes are selected whose positive ratio is higher than 1% or if they are of great importance to practical surveillance systems. The RAP dataset is the only dataset that provides viewpoint labels for each image.

The RAP-2.0 dataset is an extended version of the RAP dataset. It contains 84,928 images of 2589 different persons with resolutions ranging from 33 × 81 to 415 × 583. The attribute annotations are the same as on the RAP dataset and person identities are additionally annotated. 50,957 image are set aside for training, 16,986 for validation, and 16,985 for testing.

The PA-100K dataset consists of image captured by 598 outdoor surveillance cameras. It contains 100,000 person images with resolutions ranging from 50 × 100 to 758 × 454. The dataset is annotated with 26 binary attributes. It is randomly split into at a ratio of 8:1:1 into training, validation, and test data.

The WIDER dataset comes from the 50,574 images that usually contain many people and a lot of visual variation. A total of 13,789 images were selected. Each person is labeled with 14 distinct attributes, for the 57,524 bounding boxes. resulting in a total of 805,336 labels. This dataset was split into 5,509 training, 1,362 validation and 6,918 test images. Following the established evaluation protocol, all the 14 annotated human attributes are used. All the

person bounding boxes are cropped out, which results in 28,340 person images for training and validation and 29,177 images for testing. The un- specified labels of the WIDER dataset are treated as negative during training and excluded from evaluation.

Comparison to State-of-the-Art: A comparison of attribute recognition accuracies across several datasets is given in Table 4.1. The PGA model achieves state-of-the-art results on two datasets (PETA, PA-100K) and very competitive scores on all other datasets. Interestingly, on the WIDER dataset, which contains a wide variety of body poses, the proposed pose-attention method is narrowly outperformed by a self-attention model. On the PETA surveillance dataset with less pose variation, however, PGA achieves the better result. Table 4.2 shows the full set of label-based and example-based metrics on the PETA dataset. The proposed approach outperforms related works primarily on the example-based metrics. Since these metrics measure the overall consistency of a person's semantic description, this is a promising basis for use in a re-id system. Similar results can be observed on the RAP dataset, see Table 4.3. On this dataset the overall accuracies are significantly lower than on PETA, owing to the larger number of attributes and several more complex attribute types, such as actions. The PGA model performs strongly on most metrics, outperforms VeSPA consistently and outperforms all other approaches in the aggregate F1 measure.

Ablation Studies: Variations on the proposed attention architecture are evaluated in Table 4.4. Among single backbone models with just a weighted cross entropy loss, the ResNet-50 model performs best. The weighted loss improves recognition accuracy by about 1%. When the attention mechanism is exclusively applied, it results in a drop in accuracy, indicating that purely local attention information is not sufficient in all cases for robust attribute recognition. However, when combined with a global feature branch as originally proposed, the F-score increases significantly by almost 2%.

Individual Attributes: Table 4.5 shows the individual attribute scores by the PGA model on the PETA dataset. In addition, the positive-ratio for each attribute is shown to give an impression of the degree of imbalance for the learning task. For example, attributes like Age16-30 are almost balanced in

Table 4.1: Comparison of the PGA model to the state-of-the-art across several public datasets using F1 score.

Method	PA-100K	RAP2	RAP	WIDER	PETA
ACN [Sud15]					82.64
DeepMAR [Li15]					83.41
GoogleNet [Sze15]					84.37
WPAL-FSPP [Yu16]			43.40		83.40
WPAL-GMP [Yu16]			66.12		84.90
SRN [Zhu17, Sar18a]				85.1	84.92
HydraPlus[Liu17b]	82.53		78.05		84.07
DIL[Sar18a]				**86.4**	86.46
RAP2 [Li19]		**78.26**			
VeSPA [Sar17b]			79.59	82.4	85.49
PGA	**86.95**	78.15	**79.95**	86.2	**86.75**

Table 4.2: Additional metrics on the PETA dataset. The proposed PGA model outperforms recent approaches on all example-based metrics.

Method	mA	Acc	Prec	Rec	F1
ACN [Sud15]	81.15	73.66	84.06	81.26	82.64
DeepMAR [Li15]	82.89	75.07	83.68	83.14	83.41
GoogleNet [Sze15]	81.98	76.06	84.78	83.97	84.37
WPAL-FSPP [Yu16]	84.16	74.62	82.66	85.16	83.40
WPAL-GMP [Yu16]	**85.50**	76.98	84.07	85.78	84.90
SRN [Zhu17, Sar18a]	82.36	75.69	85.25	84.59	84.92
DIL[Sar18a]	84.59	78.56	86.79	86.12	86.46
VeSPA [Sar17b]	82.15	77.21	86.82	83.82	85.49
PGA	84.15	**79.31**	**87.80**	**85.73**	**86.75**

their relative frequency, while other attributes, such as Sandals occur only very rarely. However, the overall accuracy across all attributes is quite stable with very few outliers.

In summary, the proposed pose-guided attribute attention network achieves robust attribute reduction scores across several datasets. While the attention branch on its own does not achieve a stable result, the common combination

Table 4.3: Additional metrics on the RAP dataset. The proposed PGA model outperforms recent approaches on most example-based metrics. VeSPA is outperformed consistently.

Method	mA	Acc	Prec	Rec	F1
ACN [Sud15]	69.66	62.61	**80.12**	72.26	75.98
DeepMAR [Li15]	73.79	62.02	74.92	76.21	75.56
WPAL-FSPP [Yu16]	79.48	53.30	60.82	78.80	68.65
WPAL-GMP [Yu16]	**81.25**	50.30	57.17	78.39	66.12
VeSPA [Sar17b]	77.70	67.35	79.51	79.67	79.59
PGA	79.12	**68.15**	79.73	**80.17**	**79.95**

Table 4.4: Ablation studies for the different components of PGA on the PETA dataset. The ResNet-50 backbone with weighted loss, attention branch, and global information achieves best results.

Backbone	Weighted Loss	Attention	Att.+Global	F1
ResNet-50	✓			84.91
DenseNet-121	✓			83.23
ResNet-101	✓			84.86
ResNet-50				84.07
ResNet-50	✓	✓		83.73
ResNet-50	✓	✓	✓	**86.75**

with global features leads to an accuracy that is state-of-the-art on several datasets. Due to its diversity and the high achieved attribute scores, the PETA dataset was identified as a source for attributes that are included in subsequent re-id methods.

Table 4.5: Average accuracy of the proposed model on the PETA dataset for each of the 35 considered attributes. Additionally, the positive-ratio for each attribute is shown.

Attribute	Pos. Ratio	Accuracy
Age16-30	0.502	88.1
Age31-45	0.328	83.9
Age46-60	0.102	82.1
AgeAbove61	0.060	94.2
Backpack	0.199	88.6
CarryingOther	0.201	79.5
Causual lower	0.864	86.1
Casual upper	0.856	83.9
Formal lower	0.134	86.1
Formal upper	0.130	88.1
Hat	0.101	92.3
Jacket	0.069	77.3
Jeans	0.314	88.5
Leather shoes	0.293	89.3
Logo	0.04	73.4
Long hair	0.235	94.2
Male	0.547	93.7
MessengerBag	0.297	86.8
Muffler	0.085	95.3
No accessory	0.752	87.3
No carrying	0.272	86.8
Plaid	0.028	87.3
Plastic bag	0.075	89.0
Sandals	0.019	67.2
Shoes	0.363	81.7
Shorts	0.036	84.7
ShortSleeve	0.143	89.1
Skirt	0.045	84.8
Sneaker	0.220	82.4
Stripes	0.018	70.7
Sunglasses	0.029	69.9
Trousers	0.513	87.2
Tshirt	0.084	84.0
UpperOther	0.461	87.1
V-Neck	0.012	60.9

Chapter 5

Attribute and Pose Sensitive Re-Identification

One of the two main ideas of this thesis, as outlined in Chapter 3, is the inclusion of auxiliary information into the re-id models in order to either improve robustness and accuracy of the resulting embeddings or provide additional information to the user of a practical re-id system. In this chapter, methods to include attribute information and pose information are proposed. Initially, in Sections 5.1 and 5.2 two options for including either type of information into a CNN re-id model will be described. Then, in Section 5.3, a joint architecture, which leverages pose as well as attribute information, is proposed.

5.1 Learning Attribute-Complementary Information

Attributes represent a high level semantic description of a person. If they can be detected correctly, they describe important local details, which are often relevant to a person's identity. Attributes also provide a degree of invariance to many influences, such as pose and camera angles, which strongly influence

Figure 5.1: Examples of persons with very similar global appearance but different attributes. Attribute information can be a decisive clue to distinguish people in such cases.

conventional re-id features. Due to the localized nature of many attributes within person images, an attribute description often contains information that is missed by re-id features that focus on global appearance. Several example cases in which attribute information can help distinguish between persons whose overall visual appearance is otherwise very similar are depicted in Figure 5.1. Finally, and uniquely, compared to all other types of re-id features, attributes can also be directly communicated to human operators of a practical system. This allows for a broader range of applications and workflows, such as text based re-id queries or improved understanding of the resulting ranked lists.

5.1.1 Attribute-Complementary Triplet Loss

There are several challenges involved when using automatically detected attributes for re-id, which must be addressed:

- *Attribute disciminativeness* for the re-id task can vary strongly. For example, attributes, which are present for all persons in the data, do not add any meaningful information, based on which the re-id method can match or distinguish between persons.
- *Attribute consistency* among images of the same person recorded at different times is similarly important for re-id. If an attribute can frequently change over time and is thus not consistent across different images of the same person, it has a negative influence on the re-id result.
- *Attribute recognition accuracy* is the most important aspect. Since attribute recognition models require data with extensive attribute annotations for training, they must generally be trained on separate, specially prepared datasets. When transferring these models to new data for use in re-id, a drop in recognition accuracy will occur. Incorrectly assigned attributes will then disturb the re-id process further.

All these factors make attributes an unstable feature for re-id. Direct fusion of attributes and other re-id features can often lead to only very little or no improvement in re-id accuracy. To address these issues, this section describes a deep learning re-id approach, which includes such attribute information into the learning process of a CNN. The approach is designed with two goals in mind. Inclusion of attribute information in the training process allows the learned feature embedding to adapt to the characteristics of the attribute descriptions. It is thus possible to decrease the influence of attributes, which consistently provide unreliable information and furthermore, the embedding can focus its learning process on those cases, in which the attribute description is insufficient for re-id. The approach described in this section was previously published in [Sch17b].

The triplet loss is chosen as an objective function for the approach, since it directly models the difference between person images and thus allows for a more convenient way to incorporate added information into the distance computation than a softmax classification loss. Training samples are served to the network in sets of three: one anchor image, one match to the anchor (positive) and one mismatch (negative). The standard triplet loss is computed as

$$\begin{aligned} L_{triplet} &= \frac{1}{N}\sum_{i=1}^{N} d_i^{fp} - d_i^{fn} + m \\ d_i^{fp} &= \left\| f_i^a - f_i^p \right\|_2^2 \\ d_i^{fn} &= \left\| f_i^a - f_i^n \right\|_2^2 . \end{aligned} \tag{5.1}$$

Here, f represents the current projection from image space to embedding space as learned by the network. d_i^{fp} denotes the distance from the anchor to the positive in the learned embedding space and d_i^{fn} the distance from the anchor to the negative. Minimization of this loss encourages the feature distances between images of different persons to be large and distances for matching persons to be small. The margin m controls the strength by which these distances are separated.

Due to the direct distance computation in the loss function, an attribute vector can be integrated into the triplet training process at loss level. This is achieved by adding the attribute distances of the anchor sample to the positive and negative samples in an analogous manner. This Attribute-Complementary

Re-Id (ACR)) loss is then given as

$$
\begin{aligned}
L_{ACR} &= \frac{1}{N}\sum_{i=1}^{N} d_i^{f^p} - d_i^{f^n} + m + \gamma(d_i^{att^p} - d_i^{att^n}) \\
d_i^{att^p} &= \left\| att_i^a - att_i^p \right\|_2^2 \\
d_i^{att^n} &= \left\| att_i^a - att_i^n \right\|_2^2
\end{aligned}
\tag{5.2}
$$

where $d_i^{att^p}$ and $d_i^{att^n}$ mark the distances of the samples in triplet i based on their attribute representations. The addition of this attribute information has no direct impact on the triplet losses' gradient formulas. Take, for example, the gradient for the positive sample in the triplet:

$$
\begin{aligned}
\frac{\partial L_i}{\partial f_i^p} &= \begin{cases} 2(f_i^a - f_i^p), & if\ d_i^{f^n} - d_i^{f^p} \leq \tilde{m} \\ 0, & otherwise \end{cases} \\
\tilde{m} &= m + \gamma(d_p^{a^i} - d_n^{a^i}).
\end{aligned}
\tag{5.3}
$$

Instead, the attribute distance can be interpreted as a dynamic modification of the margin m on a triplet-to-triplet basis. Since only triplets, which violate the margin between the two distances are actually used for back-propagation through the network, this modification of the margin directly influences the cases in which the gradient is passed through the main network. If, for example, the attribute representation fails to produce a distance $d_i^{att^p} < d_i^{att^n}$, then the attribute part of the loss adds to the margin, the loss becomes more strict, and the gradient is more likely to be non-zero for this sample. Conversely, if the attribute information already achieves a good separation $d_i^{att^p} < d_i^{att^n}$, then the gradient for the learned embedding is more likely to be zero. This allows the CNN to focus on cases where the attribute information does not suffice for a successful re-id. In a sense the net learns to re-rank a basic ranking generated by the attributes. The approach is motivated by the assumption that this might be a simpler task than re-id without attribute information. A parameter γ is used to control the degree of influence of the attribute information on the loss.

While the described approach allows the network to learn a feature embedding that is complementary to the pre-existing attribute feature, it may still be influenced negatively by unreliable attribute information. Thus, an additional layer for weighting the attributes during training is introduced. The weight layer simply performs an elementwise multiplication of the attributes with a learned global weight vector.

At test time, attribute information is first used to compute an initial attribute distance between person images. The complementary CNN feature distance is then combined with the weighted attribute just as during training.

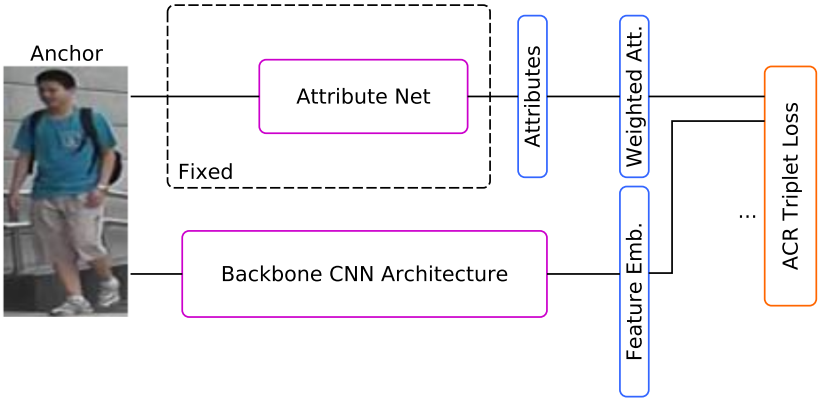

Figure 5.2: Configuration of the proposed triplet loss architecture. For visual clarity only the anchor branch of the network is displayed.

5.1.2 Evaluation

The proposed ACR triplet loss for learning attribute-complementary information is evaluated using attribute scores provided by the pose-attention attribute model described in Chapter 4. Attributes are learned on the PETA dataset, as this dataset provides the largest degree of visual variation and the

resulting attributes are thus more likely to generalize to the re-id datasets used for this evaluation. The evaluation will focus on the degree to which ACR can leverage attribute information in relation to standard feature fusion techniques and other baselines. Then, the learned attribute weights are analyzed in order to gain insight into which attributes are useful to re-id and which are not. Finally, an alternative method of leveraging information from the attribute recognition network is evaluated.

Datasets: The ACR model will be evaluated on the two most popular re-id datasets, Market-1501 and DukeMTMC-reID. The Market-1501 dataset [Zhe15b] provides 32,668 cropped images of 1,501 persons, which were generated by automatic person detection in 6 cameras. 751 persons are used for training. For testing, a set of 3,368 query images is available. The gallery size of the Market-1501 dataset is 19,734 and contains 2,793 distractors. The DukeMTMC-reID dataset [Zhe17b] consists of persons cropped from the DukeMTMC tracking dataset [Ris16], which is recorded by 8 cameras. The dataset consists of 1,812 different persons of which 1,404 appear in more than one camera. 702 persons are set aside for the training set and the remaining 1,110 are used for testing. This results in a training set of 16,522 images, a probe set of 2,228 images and a gallery set of 17,661 images. The performance will be evaluated based on mAP and Rank-1, -5, -10 and -20 accuracies to give an impression of the CMC. On both Market-1501 and DukeMTMC-reID the provided evaluation code is used.

Baselines: Four simple baseline methods are defined to compare the proposed ACR model to:

- **ReIdCNN**: This baseline trains a plain CNN model of the same architecture as the ACR model but without any attribute information. A standard triplet loss and the same training settings as for ACR are used.
- **Attributes**: For this baseline, the direct performance of the attribute scores generated by the attribute net is evaluated on the target re-id dataset. No learning on the target data is involved in this method.

- **Attributes-KISSME**: This baseline applies KISSME metric learning [Koe12] using the attribute predictions on the target data. This baseline indicates the potential for re-id contained in the attribute predictions.
- **ReIdCNN+Attributes**: In order to show the complementary nature of the information learned by the ACR model, this baseline performs a simple score fusion between the ReIdCNN and Attributes-KISSME baselines. Similar to ACR, attributes scores are weighted with 0.5 (the value of γ in ACR).

The results of our baselines are given in Tables 5.1 and 5.2 for Market-1501, and DukeMTMC-reID, respectively. All models use the ResNet-50 architecture as a backbone. On both datasets similar trends can be observed. The ReIdCNN baseline performs strongly while pure attribute information can only achieve a very low person re-id accuracy. The main reasons for this are the comparatively low dimensionality of the attribute information and their presumably limited reliability due to varying performance of the attribute classifier. However, the application of KISSME metric learning to the attribute predictions on the target dataset shows that a higher potential for re-id is contained in the predictions. This indicates that certain attributes are helpful for re-id while others distort the result without metric learning. Finally, the combination of the ReIdCNN with attribute information yields the best baseline performance but the result is dominated by the CNN and the overall improvement through attributes is very slight.

Table 5.1: Results of the ACR approach on the Market-1501 dataset.

Method	mAP	r1	r5	r10	r20
ReIdCNN	59.75	79.67	90.25	93.10	96.43
Attributes	8.39	14.61	25.72	46.11	59.98
Attributes-KISSME	17.91	26.24	43.18	53.27	61.77
ReIdCNN+Attributes	60.45	**80.13**	91.56	94.23	96.51
ACR	**61.25**	79.17	**92.43**	**95.39**	**97.01**

Table 5.2: Results of the ACR approach on the DukeMTMC-reID dataset.

Method	mAP	r1	r5	r10	r20
ReIdNet	48.76	66.78	79.15	85.46	88.52
Attributes	7.45	11.98	23.12	31.67	42.45
Attributes+KISSME	13.04	20.56	45.03	53.21	65.71
ReIdCNN+Attributes	50.41	69.34	80.28	87.13	90.78
ACR	**51.38**	**71.14**	**83.68**	**89.44**	**90.00**

ACR Results: When combining the attribute information with CNN features through the proposed ACR triplet loss, another significant boost in re-id accuracy is achieved. Compared to the ReIdCNN baseline, the use of attribute information in ACR can improve the resulting performance by 1.50% mAP on Market-1501 and 2.62% mAP on DukeMTMC-reID. Compared to the ReIdCNN+Attributes baseline score fusion, the attribute information is much better exploited. The ACR CNN did not have to learn the information, which is already contained in the attributes and thus could use more of its parameters for learning to compensate in failure cases. The ReIdCNN+Attributes baseline in contrary contains redundant information in the CNN, which leads to a reduced benefit of attributes. Qualitative results of the ACR model are depicted in Figure 5.3.

However, due to their low dimensionality attribute vectors carry only a limited amount of discriminative information. In order to evaluate the potential of using other components of the attribute recognition network, the earlier *pool-5* layer in the network is used as a feature extractor. The extracted embedding's performance for re-id on the Market-1501 dataset, as well as its impact on ACR is evaluated in Table 5.3. It can be seen that the embedding achieves a much higher standalone accuracy and also has a beneficial impact on ACR. Thus, using earlier information with less explicit semantics from within an attribute network can provide more relevant information for re-id.

Attribute Weighting: Table 5.4 shows the attributes that are most and least strongly weighted by the approach on the Market-1501 dataset. There is a

Table 5.3: ACR results when based on attributes or an embedding vector from the attribute net on the Market-1501 dataset.

Method	mAP	r1	r5	r10	r20
Attributes	8.39	14.61	25.72	46.11	59.98
Embedding	19.40	41.29	62.28	70.19	78.41
ACR (Attributes)	61.25	79.17	92.43	95.10	97.01
ACR (pool-5)	**65.19**	**83.61**	**94.27**	**95.53**	**97.62**

clear correlation between the attribute weighting and their original accuracy on the source dataset. Many of the highly rated attributes are common in the Market-1501. Furthermore, many of them focus on larger parts of a person, such as upper or lower body, which is less difficult to locate than attributes with very small spatial occurrence. Unsurprisingly, the attributes rated lowest by ACR include those that occur rarely and are very specific. Unfortunately, it is exactly such rarely occurring attributes which are most discriminative for re-id of those individuals that have them.

Table 5.4: Attributes from the PETA dataset that were considered most and least relevant for re-id on the Market-1501 dataset.

Most relevant	Least relevant
upperBodyLongSleeve	accessoryKerchief
accessoryNothing	footwearPurple
hairShort	accessoryFaceMask
hairBlack	accessoryShawl
footwearBlack	lowerBodyPink
personalLess30	hairPurple
lowerBodyBlack	lowerBodyLogo
lowerBodyTrousers	hairRed
upperBodyBlack	hairGreen
personalFemale	hairOrange

It can generally be observed that attributes which are visible only in very small portions of a person image are determined to be of less help for re-id by ACR. Similarly, attributes which are presumably very rare or not at all present on the target dataset, receive a low weight (e.g. orange hair).

In summary, the proposed ACR model better exploits attribute information than other established feature fusion methods. ACR additionally provides insight into which attributes are relevant to re-id and which are not. Overall, however, the resulting accuracy is dominated by the learned feature as attributes provide only a small increase in accuracy. A promising alternative is the extraction of further information from earlier in the attribute network. While semantics are not explicit here, a finer degree of information is available and potentially misclassified attributes have no negative impact.

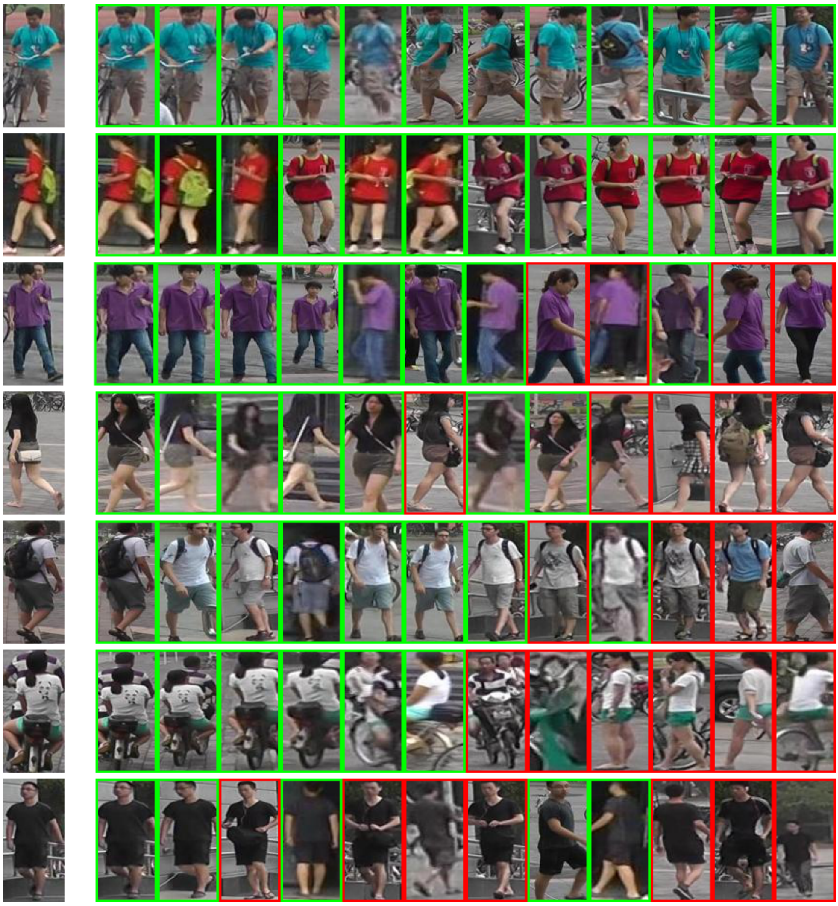

Figure 5.3: Qualitative results of the ACR approach for challenging queries on the Market-1501 dataset. The query images are displayed on the left and the top 12 results are shown. Correct matches are highlighted green and false matches in red. Note that false results are often semantically and visually similar to the query.

5.2 A Pose-Sensitive Embedding

One of the main factors determining the visual appearance of a person in surveillance images and video is their body pose and relative orientation to the camera. Furthermore, across cameras in a camera network, pose and view are usually much more strongly influencing the variation in appearance of the same person than other factors, such as image quality. Thus, explicitly including this information into the learning process of a re-id model is expected to be beneficial to the resulting re-id accuracy.

A number of works already exist that follows this general motivation, see Section 2.1.1.5. However, such works have relied on either fine-grained pose information (e.g., joint keypoints) or coarse information (e.g., orientation to the camera). In contrast, this section describes a proposed model, which includes both these levels of granularity into a Pose-Sensitive Embedding (PSE). Coarse view is used to learn a set of embeddings specialized on a coarse set of views. This can help the network to better address the variation in appearance between different view angles (see also Figure 5.4). Finer-grained pose information in the form of keypoints is included at the beginning of the network to aid the network in localizing relevant body regions and possibly even identify and compensate for errors resulting from a previous detection stage (see Figure 5.5). The models described in this section were developed in joint previously published works and with equal contribution of Saquib Sarfraz [Sar17b, Sar18b]

5.2.1 View Information

For encoding a coarse view information indicating a person's orientation relative to the camera, the quantization ['front', 'back', 'side'] is used. This view information is then included into the re-id network through a view prediction side-branch. The branch is based on a common trunk with the main re-id stream of the network and learns to recognize the view of the person depicted in the image through a 3-class softmax cross-entropy loss (see Figure 5.6). The

Figure 5.4: Different views or orientations of the same person. Depending on the angle, the visual appearance can vary significantly. In the given example, the black backpack can have a strong influence on the upper body color depending on the orientation.

Figure 5.5: Person detections and corresponding estimates of joint locations. The pose information indicates the different locations of the same body parts between images. Pose information can also help in indicating detection error (i.e. mis-aligned bounding box in the lower right example).

prediction scores are then used in the main re-id network to train three feature embeddings, each of which is specialized to encode information specific to one of the views. For this, the main architecture is branched into three separate units, whose output feature embedding is multiplied with the prediction score of one of the three views. This weighting of embeddings with

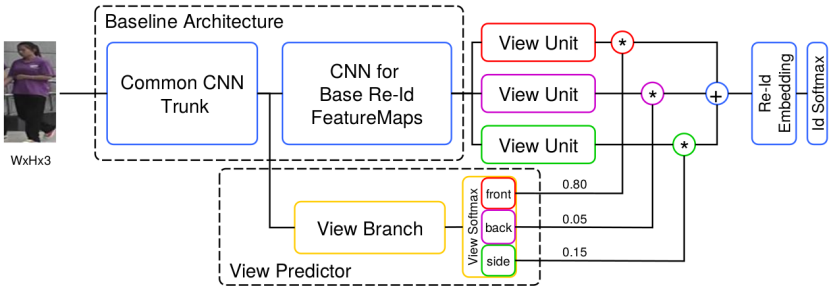

Figure 5.6: View-prediction branch and view units. Multiplication of the unit's output embedding with the corresponding view prediction score leads to the unit specializing in encoding information from that specific view. The view branch is pre-trained on separate data.

the corresponding view modulates the gradient flowing through the units. For example, a training sample with a strong 'front'-view prediction will result in a large weight for the embedding multiplied with the front-score and very small weights for the two other embeddings. As the weight persists in the gradient of the multiplication layer, the gradient reaching the two other embeddings will be low, while the front-embedding will receive an almost undiminished gradient and learn to adapt further based on the current frontal image. This procedure allows each unit to learn a feature map specialized for one of the three views.

An important requirement of the approach is the availability of suitable view annotations to train the view prediction branch. It cannot generally be assumed that view annotations are available on the re-id dataset on which the embeddings are trained. Thus, pretraining of the view classifier is carried out on the separate RAP [Li16a] pedestrian attribute dataset, which provides such annotations. The resulting classifier is then directly transferred to the re-id model. On the re-id dataset training of the main re-id model takes place while the weights of the view prediction branch remain fixed. The common trunk is fine-tuned at low learning rate. Due to the low-level nature of the feature representations, the output of the view branch is not strongly impacted by this.

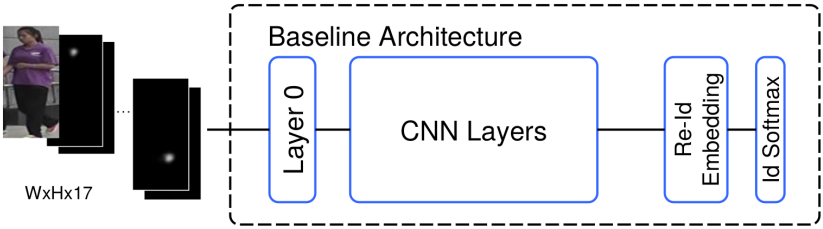

Figure 5.7: Pose information can be included into a re-id CNN by adding additional channels – one for each body joint location. If the network is initialized with standard ImageNet weights, the first layer of the network will have to be adapted for the new input size.

5.2.2 Full Body Pose

Person pose estimation methods typically provide a list of body joint locations in the image as their output. For example, the off-the-shelf DeeperCut model [Ins16] provides 14 such keypoints. While such pose information has been used for re-id previously, it is typically used to either normalize the input image or define a set of local regions that are used to integrate local information into the re-id embedding (Section 2.1.1.5). In contrast to this, the proposed method of integrating the joint location information is to add a new input channel for each of the 14 keypoints (see Figure 5.7). Each channel then represents a location probability map for that given body joint. The intention behind providing pose information in this way as part of the input is twofold:

- On a global scale, the set of joint locations provide a rough segmentation of the person depicted in the image. This information can be used by the network to prevent background information from diluting the resulting re-id embedding.
- At a local level, individual keypoints or pairs of keypoints provide the location of body parts, which can help the network to better learn a pose independent feature embedding.

The body joint channels thus serve as a simple attention mechanism at the earliest possible stage. It is left to the network to learn how to best apply

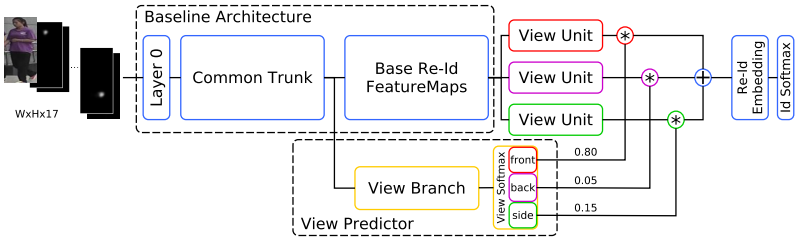

Figure 5.8: The combined pose-sensitive embedding (PSE) model. Body keypoint information is provided as part of the input and a view prediction branch helps to learn three specialized embeddings, which are combined into a final, pose-sensitive embedding of re-id.

the body joint information into the resulting embedding. This flexibility can be further increased by not providing the final joint location decisions of the pose estimator, but rather the internal probability map based on which the locations were determined. This allows the re-id network to take into account the certainty with which a certain joint point is located. It may even help identify cases where high uncertainty may have led to a faulty pose estimate, which should not be relied upon for re-id.

5.2.3 Full PSE Model & Staged Training

The two described methods for including view and pose information do not only target different granularities but are intentionally integrated at very different locations of the re-id network. They are thus easy to combine into a joint model, which is depicted in Figure 5.8.

Depending on the type of pose information included, the training procedure of the model requires different stages. For training each model, the backbone CNN is initialized with weights pretrained for ImageNet classification. In order to train a model with view information, the view-predictor branch is first fine-tuned on the RAP dataset [Li16a]. Then, only the specialized view

units and the final person identity classification layer are trained on the target re-id dataset, as these layers are initialized with random weights. When training an embedding including body keypoint input, the ImageNet weights do not match the size of the increased number of input channels. To adapt the network for the larger input, the initial convolutional layer and the final ID classification layers are thus trained from scratch while the remaining weights stay fixed. Once these two layers are adapted to the remainder of the network, the full network can be fine-tuned for re-id. When the joint PSE model is trained, the pose information is first provided so that the view predictor branch can already benefit from it while it is fine-tuned on the RAP dataset in conjunction with the common trunk segment of the main network.

All described CNN embeddings are trained using the same protocol. The input images are normalized to channel-wise zero-mean and a standard variation of 1. Data augmentation is performed by resizing images to 105% width and 110% height and randomly cropping the training sample, as well as randomly flipping it horizontally. Training is performed using the Adam optimizer at default parameters with an initial learning rate of 0.0001 and a decay of 0.96 every epoch.

5.2.4 Evaluation

The proposed pose-sensitive embedding (PSE) will be evaluated on the Market-1501 and DukeMTMC-reID datasets, as well. Similar to ACR, the primary backbone architecture will be ResNet-50, unless otherwise stated. Evaluation is carried out using the mAP and rank accuracy metrics and the official evaluation code is used. The evaluation will first focus on the influence of the different types of pose information on the re-id embeddings. Two different baseline architectures are investigated and the semantic content of the proposed view branch on the target dataset will be analyzed qualitatively.

Type of Pose Information: The impact of the different types of pose information on the re-id embedding is summarized in Table 5.5. Experiments are carried out across both datasets and using two types of backbone architectures to demonstrate the stability of the results. When the view predictor

is used, it is branched off the main network after the Reduction-A block for the Inception-v4 architecture and view units are added by replicating the final Inception-C block three times. In case of the ResNet-50, the branch-off occurs after convolutional block 2 and replication is applied for convolutional block 5.

Table 5.5: Impact of different types of pose information (i.e. view and keypoint body pose) on the re-id embedding across two datasets and network architectures.

CNN	Method	Market-1501					DukeMTMC-reID				
		mAP	**r1**	**r5**	**r10**	**r50**	**mAP**	**r1**	**r5**	**r10**	**r50**
Inception-v4	Baseline	51.9	75.9	89.8	92.5	97.3	36.6	61.8	74.8	79.8	89.4
	Views only	61.9	81.5	92.3	94.9	98.1	40.3	62.7	76.6	81.1	90.3
	Pose only	60.9	81.7	91.8	94.4	97.9	48.2	70.5	81.9	86.1	92.7
	PSE	**64.9**	**84.4**	**93.1**	**95.2**	**98.4**	**50.4**	**71.7**	**83.5**	**87.1**	**93.1**
ResNet-50	Baseline	59.8	82.6	92.4	94.9	98.2	50.3	71.5	83.1	87.0	94.1
	Views only	66.9	**88.2**	**95.4**	**97.2**	98.9	56.7	76.9	87.3	90.7	95.7
	Pose only	61.6	82.8	93.1	95.5	98.3	53.1	73.4	84.5	88.1	94.3
	PSE	**69.0**	87.7	94.5	96.8	**99.0**	**62.0**	**79.8**	**89.7**	**92.2**	**96.3**

It can be seen that inclusion of either view or fine-grained pose information leads to significant improvements in accuracy of re-id. The Inception-v4 architecture achieves the overall smaller accuracy but receives a larger absolute improvement by including pose information. Inclusion of views into the ResNet-50 model provides consistent improvements of 7.1% mAP and 5.6% rank-1 on Market as well as 6.4% mAP and 5.4% rank-1 for Duke. The average improvement resulting from the inclusion of pose is a little smaller, 2-3% in mAP.

Interestingly, a combination of both types of information leads to a clear further improvement in accuracy. For ResNet-50, mAP is increased by 2.1% on Market and 5.3% on Duke in reference to the best result of either views or pose. A similar observation can be made for the Inception-v4 model. This further increase in accuracy clearly suggests some degree of complementary information between the two options. However, since the joint locations do implicitly contain view information as well, it stands to reason that the way in which the information is integrated is equally important as the potentially

complementary nature. In this case, it is likely that the inclusion of pose information at such different positions in the network, and thus different degrees of spatial resolution and semantic content, has an equally large impact on the observed increase in accuracy.

Analysis of the View Branch: Since view annotations are not readily available on either Market-1501 or DukeMTMC-reID, the view branch was pretrained on the RAP dataset for all experiments. On RAP the branch achieves robust accuracies of 82.2%, 86.9% and 81.9% after training for front, back, and side view, respectively. How much this accuracy degrades when the branch is applied on the two target re-id datasets cannot be quantitatively assessed, due to lack of labels. However, a qualitative insight into the average prediction accuracy can be gained by performing a view prediction on the entire target dataset test splits and averaging the images for each of the three view prediction values. Such mean images for all three datasets, i.e. Market-1501, DukeMTMC-reID, and RAP, are depicted in Figure 5.10.

On all three datasets some clear differences are visible between the front and back mean images. Particularly in the front mean image a clear face region as indicated by a lighter, skin-colored blob can be made out. Conversely, the back view does not have such a region. The side view is less clear. The less distinct area of the legs results from the much larger variety in leg positions in side view images. Furthermore, the left and right side are modeled as one class and thus lead to an additionally stronger blurring when combined in the mean image. And lastly, the view annotations on the RAP dataset assign a fairly narrow angle range for frontal and back view while everything else is considered side view. Thus, this class is expected to result in stronger blur just by the annotation strategy alone.

A clearer impression of the view prediction quality can be gained when the RAP mean images based on predicted values are compared to mean images based on annotations in Figure 5.11. The images look almost identical, and particularly the side image shows similar effects in both cases. The overall comparison of the RAP mean images to the mean images of the other two datasets give a positive qualitative indication of view predictor accuracy on the target datasets.

Comparison to Explicit Pose Use: In a recent publication, the Pose Invariant Embedding (PIE) [Zhe17a] was proposed as a more explicit way of including pose into a CNN-based re-id model. In contrast to the PSE method, PIE uses the estimated pose information to explicitly align body parts by generating a *PoseBox* image. The re-id network is then provided with the original image, the PoseBox image and the pose estimator's confidence score. While PSE leaves it to the network to learn how exactly to use the pose information in its feature map, providing such pose-normalized data is an interesting alternative.

Table 5.6 gives a comparison between PIE and PSE on the Market-1501 dataset, using the same ResNet-50 backbone architecture. The PSE model clearly outperforms the PIE model by 15.1% in mAP and 9.0% in rank-1. Part of this is the better trained PSE baseline model but even when the absolute increase over the baseline model is considered, the increase of PSE over its baseline is significantly larger.

Table 5.6: Comparison between explicit use of pose in the Pose Invariant Embedding (PIE) [Zhe17a] and the proposed Pose Sensitive Embedding (PSE).

	Method	mAP	R-1	R-5	R-10
PIE	Baseline1 (R,Pool5)	47.6	73.0	87.4	91.2
	PIE (R,Pool5)	53.9	78.7	90.3	93.6
	Difference to Baseline	6.3	**5.7**	**2.9**	**2.4**
PSE	Baseline (Resnet-50)	59.8	82.6	92.4	94.9
	PSE (Resnet-50)	69.0	87.7	94.5	96.8
	Difference to Baseline	**9.2**	5.1	2.1	1.9

In summary, the proposed pose-sensitive embedding achieves a strong increase in re-id accuracy. Both components contribute strongly and further increase accuracy, when combined with each other. Besides potential complementary information, a main reason for the increased accuracy of the combined architecture is likely the different positions and levels of semantics at which each information is included.

Figure 5.9: Qualitative results on the Market-1501 dataset. Two results are shown for each query. The upper result corresponds to the model without inclusion of pose information. The lower result is obtained by inclusion of pose information.

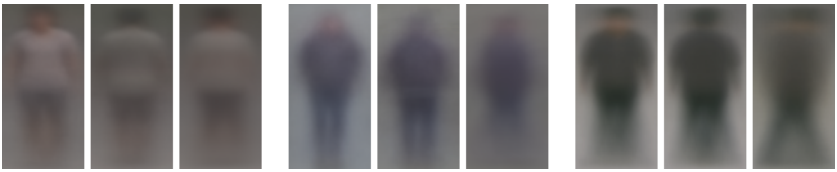

Figure 5.10: View prediction mean images of Market-1501 (left), DukeMTMC-reID (middle) and RAP (right)

Figure 5.11: Comparison of mean view images from RAP, which were generated by relying on the view predictions (left) and ground truth view annotations (right).

5.3 A Pose-Sensitive Attribute-Attention Model

During development of the models described in the previous two sections, several observations were made:

- While attribute information can be of help in re-id, its benefit is mitigated by the fact, that attribute prediction vectors have a low dimensionality and thus contain only a limited degree of information. Furthermore, if an attribute prediction is wrong, the entire information (i.e. dimension) is of little use and may actually be counter-productive to re-id. On the other hand, results have shown that feature embedding from previous stages in the attribute recognition network can be of notably more help in re-id.
- Pose and view information were shown to both boost re-id accuracy when used within the same network. However, the seemingly complementary nature of the two types of information is mitigated by the fact that pose keypoints do in fact already contain an encoding of the depicted view. For example, the relative position between left and right shoulder joint gives a strong clue towards the person's orientation to the camera. Thus, it seems likely that the method by which and particularly the different locations at which the two types of information were integrated into the network played a key role in improving accuracy.

Building on these observations, a new joint model is proposed, which includes information from the described attribute network, as well as pose information. Note that the view branch is already a method of including a single attribute, i.e. view, into the pose-sensitive re-id model. However, this technique can not easily be extended to further attributes, as the number of required specialized units would be very high and add many parameters to the network. Furthermore, few other attributes are suited to this type of integration, as they are highly unbalanced and often much more localized in the image than the visual cues based on which the view can be estimated. Thus, the attribute branch is

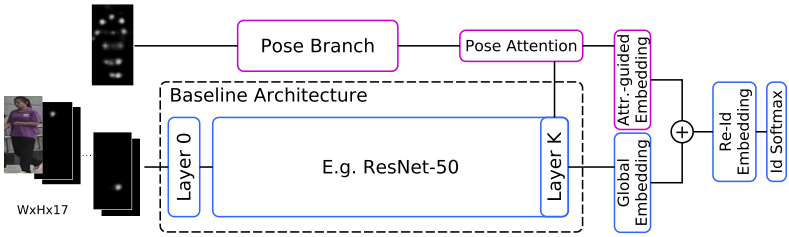

Figure 5.12: The proposed final model for combination of pose information as input channels and attribute-guided attention maps computed from a pose representation. The upper pose branch and the pose-guided attribute attention maps are directly taken from the PGA model, see Figure 4.2, and remain fixed during training of this model.

removed in the joint model and replaced with the pose-guided attribute attention branch learned in the attribute recognition network. This modification has several advantages:

- Compared to direct use of attributes, the attribute attention maps guide the re-id network to focus on certain spatial regions, but leave the network free to learn the type of information that will be extracted at that location. Thus, the attention maps for all attributes can be included and no mechanism is required to filter out misleading information, such as wrongly predicted attributes.
- In comparison to the view prediction branch, the attribute attention maps do not require many added parameters and scale better, if the number of attributes increases.
- Lastly, the attribute attention masks are learned directly from pose information. As pose is an intermediate level of semantics, the attribute masks are likely to transfer better to different datasets. As long as the pose estimator functions reliably, the attention maps will not be impacted by changing characteristics in the images.

The final model is depicted in Figure 5.12. The training process is similar to that of the previously described PSE embedding, except that no common

network trunk between the re-id model and the attention branch complicate the training procedure. The attention branch, i.e. the pose branch and the attribute attention maps, is pre-trained on an attribute dataset and remains fixed during training of the re-id model. Thus, the gradient of the re-id softmax cross-entropy loss flows through the attribute-guided embedding and the corresponding attention maps into the main backbone network where it combines with the gradient of the global branch of the architecture. In this way, the pose-guided attribute attention shows the re-id network where to look spatially, as determined by the various attributes, but leaves the network free to learn which type of information to extract at the specified spatial position. In contrast to the rigid attribute information employed in Section 5.1 this remaining degree of freedom has two advantages. Firstly, no counterproductive information, such as mis-classified attributes, need to be compensated by the resulting re-id network. Secondly, attributes are often very localized and the resulting attention maps force the re-id network to extract information from spatial regions which it otherwise might not have relied on. This prevents an overly strong focus on salient image regions, which can have a negative impact on re-id accuracy.

5.3.1 Evaluation

In this section the performance of the combined Pose-Sensitive Attribute Embedding (PSAE) will be analyzed in detail. To that end, the model will first be compared to the default PSE model, as well as recent state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets. PSAE will also be compared to PSE under several evaluation settings with more practical relevance, including video-based re-id on the MARS dataset, scalability with increased numbers of distractors on the Market-500K dataset, and its ability to work in conjunction with person detectors on the PRW dataset.

Comparison to PSE and State-of-the-Art: In Table 5.7, the attribute attention mechanism's inclusion into a re-id embedding is compared to the two types of pose information. The attribute attention outperforms both pose and view clearly and consistently across most ranks and both datasets. When the

view branch is replaced with the attention branch, as proposed in Section 5.3, the overall accuracy of the resulting PSAE network surpasses that of PSE by 4.2% and 1.2% in mAP on Market and Duke, respectively. Thus, the pose-attention branch is a more powerful addition than the view branch of the PSE model.

Table 5.7: Comparison of the attention branch to pose and view information across two datasets and network architectures.

CNN	Method	Market-1501					DukeMTMC-reID				
		mAP	R-1	R-5	R-10	R-50	mAP	R-1	R-5	R-10	R-50
ResNet-50	Baseline	59.8	82.6	92.4	94.9	98.2	50.3	71.5	83.1	87.0	94.1
	Views only	66.9	88.2	95.4	97.2	**98.9**	56.7	76.9	87.3	90.7	95.7
	Pose only	61.6	82.8	93.1	95.5	98.3	53.1	73.4	84.5	88.1	94.3
	Attn only	**67.3**	**89.3**	**96.1**	**97.7**	98.8	**57.9**	**75.1**	**84.9**	**89.3**	**96.0**
	PSE	69.0	87.7	94.5	96.8	99.0	62.0	79.8	89.7	92.2	96.3
	PSAE	**73.2**	**90.1**	**96.3**	**97.9**	**99.2**	**63.2**	**80.6**	**91.2**	**92.9**	**97.1**

In Table 5.8 the state-of-the-art is compared with the performance of the proposed PSAE and PSE embeddings on the three datasets Market-1501, Duke, and, additionally, MARS. Since the PSE model often achieves a lower accuracy compared to PSAE, the discussion of results will focus on PSAE.

The MARS dataset [Zhe16a] is based on the same raw data as the Market-1501 dataset. In contrast to Market-1501, MARS is providing tracklets of persons instead of single images. Therefore MARS is well suited to evaluate the performance of re-id approaches in video. The dataset consists of 8,298 tracklets for training and 12,180 tracklets for testing with 509,914 and 681,089 images respectively. In order to apply PSE and PSAE to MARS, for each tracklet, the mean descriptor across all images in the tracklet is computed. This straightforward extension of a single-image embedding to videos does not include any actual temporal information but is well suited to stabilize the resulting feature vector across the track and usually results in competitive accuracy.

Across all three datasets, a consistent improvement by PSAE over the ResNet-50 Baseline model and the PSE embedding is observed. Particularly on the Market-1501 dataset, a significant additional boost by 4.2% mAP and 2.4%

Table 5.8: Comparison of the proposed PSE and PSAE approaches with the published state-of-the-art. In the top section of the table, the PSE embedding is compared to state-of-the-art methods not using re-ranking. In the lower part, re-ranked results are reported using either k-reciprocal [Zho17a] or ECN re-ranking [Saq18].

Method		Market-1501		Duke		MARS	
		mAP	R-1	mAP	R-1	mAP	R-1
GAN[Zhe17b]	ICCV17	56.2	78.1	47.1	67.7	-	-
Latent Parts [Li17a]	CVPR17	57.5	80.3	-	-	56.1	71.8
ResNet+OIM [Xia17]	CVPR17	-	82.1	-	68.1	-	-
ACRN[Sch17c]	CVPR17-W	62.6	83.6	52.0	72.6	-	-
SVD [Sun17]	ICCV17	62.1	82.3	56.8	76.7	-	-
Part Aligned [Zha17c]	ICCV17	63.4	81.0	-	-	-	-
PDC [Su17a]	ICCV17	63.4	84.1	-	-	-	-
JLML [Li17d]	IJCAI17	65.5	85.1	-	-	-	-
Forest [Zho17b]	CVPR17	-	-	-	-	50.7	70.6
DGM+IDE [Ye17]	ICCV17	-	-	-	-	46.8	65.2
QMA [Liu17c]	CVPR17	-	-	-	-	51.7	73.7
MGCAM[Son18]	CVPR18	74.3	83.8	-	-	**61.3**	**75.7**
HA-CNN [Li18b]	CVPR18	76.7	91.2	63.8	80.5	-	-
Mancs[Wan18a]	ECCV18	**82.3**	**93.1**	**71.8**	**84.9**	-	-
ResNet-50 Baseline		59.8	82.6	50.3	71.5	49.5	64.5
PSE		69.0	87.7	62.0	79.8	56.9	72.1
PSAE		73.2	90.1	63.2	80.6	58.9	74.3
PSE + k-reciprocal		83.5	90.2	78.9	84.4	70.7	74.9
PSE + ECN (rank-dist)		84.0	90.3	79.8	85.2	71.8	76.7
PSAE + k-reciprocal		84.6	91.9	81.0	86.1	71.7	77.1
PSAE + ECN (rank-dist)		**84.8**	92.1	**81.2**	**86.9**	**72.5**	**78.1**

rank-1 accuracy is achieved. On all three datasets, the proposed embeddings achieve very high scores and are only outperformed by very recent new publications.

The bottom part of Table 5.8 additionally provides re-ranked results using either the k-reciprocal embedding [Zho17a] or the expanded cross neighborhood (ECN) re-ranking. On all datasets re-ranked PSAE results achieve top accuracies with ECN being slightly better than k-reciprocal re-ranking. The re-ranked embedding results surpass the state-of-the-art on Market by 2.5% in mAP, on Duke by 9.4% in mAP, and on MARS by 11.2% in mAP.

Large Gallery Sizes: For evaluation of the robustness of the proposed models in real-world deployments with very large gallery sizes, the Market-1501+500k (Market500k) dataset is used. This dataset is an extension to the Market-1501 dataset and offers an additional 500,000 distractor images that can be added to the gallery to evaluate the impact of increasing gallery size. The established protocol is to evaluate a model by stepwise adding 100k, 200k, 300k, 400k, and finally all 500k distractor images to the Market-1501 gallery. The distactors from the Market500k set are chosen randomly.

Table 5.9: Evaluation of performance drop of the proposed embeddings and related works on the Market-1501+500k distractors dataset.

Method	mAP by #Distractors				R-1 by #Distractors			
	0	100k	200k	500k	0	100k	200k	500k
I+V [Zhe18]	59.9	-12.7%	-18.0%	-24.5%	79.5	-7.2%	-10.1%	-14.1%
APR [Lin17]	62.8	-10.0%	-14.7%	-20,7%	84.0	-4.9%	-6.9%	-10.2%
TriNet [Her17]	69.1	-10.4%	-15.1%	-22.4%	84.9	-6.1%	-8.5%	-12.0%
ResNet-50 Baseline	59.8	-8.7%	-13.4%	-20.6%	82.6	-5.9%	-8.4%	-11.4%
PSE	69.0	-8.1%	-11.9%	-18.1%	87.7	-4.1%	**-5.8%**	**-7.9%**
PSAE	**73.2**	**-7.7%**	**-10.2%**	**-17.8%**	**90.1**	**-4.0%**	-6.0%	-8.3%

The performance of the proposed models under such growing gallery sizes is given in Table 5.9 and compared to the limited number of approaches that have published results on the dataset. For 0 added distractors the absolution mAP value of all models is reported. With every addition of distractors, the table then reports the relative change in mAP to indicate how well the model copes with the larger gallery. A clear difference between the proposed models and related work becomes visible. Whereas the best related method (TriNet [Her17]) is decreasing by 22.4% in mAP and 12.0% in rank-1 accuracy when adding 500k distractors, the PSAE approach only loses 17.8% mAP and 8.3% in rank-1. The PSE model shows a similar trend, with stronger drop in mAP but smaller drop in rank-1. This indicates that the PSAE model is slightly better able to find further correct matches after the first one. The strong differences

to related approaches also indicate that pose information and attribute attention help to better maintain the model's accuracy with a growing number of distractors.

Person Detection & Re-Id: Re-id approaches typically rely on pre-cropped person images provided in the dataset. Thus, if the images were not cropped by an automatic detector, the re-id approach will become biased towards more precise hand-cropped boxes. Even if a person detector was used, no flexibility in detector choice is possible. The Person Re-Identification in the Wild (PRW) dataset offers full camera images in which persons have to be found. Approaches can rely on their own solution to person detection or alternatively on automatic detections created by a DPM detector, wich are provided alongside with the dataset. Like MARS, PRW is based on the same video material as the Market-1501 dataset and provides 13,126 training, 140,469 test and 2,057 query images cut out from video streams of six cameras. While the videos used to create the dataset are the same as for Market, person identities are not consistent and there are only 482 IDs in the training and 450 IDs in the test set. Since the PRW datasets unfiltered DPM detections, false detections and misaligned images are included. Figure 5.13 shows examples of detection errors contained in the data.

Figure 5.13: Example images of the PRW test set showing false and mis-aligned person detections.

Confidence scores are provided with the detection and for re-id, a threshold has to be chosen on which detections will be used for the re-id process and which are discarded. Higher threshold values will result in cleaner detections

but also in cases where the query person is wrongfully discarded and the re-id method cannot recover this type of error. The PRW evaluation protocol specifies calculation of mAP and rank accuracies under different confidence thresholds. For this, the threshold is defined by the average number of person detections remaining in each image.

Table 5.10 compares the state-of-the-art of the PRW dataset with results of the proposed approaches for 3, 5, 10, and 20 average detections per frame.

In comparison to the state-of-the-art, both proposed approaches achieve improvements with the final PSAE embedding having a 8.2% higher mAP and 4.5% higher rank-1 score. Furthermore, both models show a very low reduction in rank-1 with increasing numbers of detections per frame, i.e., with an increasing number of false detections and badly aligned images being processed. The mAP actually increases, indicating that the model is well able to sort out false detections and reliably identify the added correct detections that occur when additional detections per image are taken into account.

In summary, the proposed combined model achieves very competitive accuracies across a number of challenging evaluation setups with real-world aspects. The PSE embedding is outperformed in most cases by the PSAE model, indicating that the attribute attention maps are a better addition to the network than a view classification branch. This is particularly notable, since the view branch and the corresponding view units increase the number of parameters in the network, which the attention maps do not.

Runtime Considerations: Runtime is an important aspect during practical use of any re-id system. While the methods described in this thesis are not

Table 5.10: Comparison of the proposed models with the state-of-the-art on the PRW dataset.

Detector	Method	#detections=3			#detections=5			#detections=10			#detections=20		
		mAP	R-1	R-20	mAP	R-1	R-20	mAP	R-1	R-20	mAP	R-1	R-20
DPM	IDE_{det} [Zhe16b]	17.2	45.9	77.9	18.8	45.9	77.4	19.2	45.7	76.0			
DPM-Alex	IDE_{det} [Zhe16b]	20.2	48.2	78.1	20.3	47.4	77.1	19.9	47.2	76.4			
DPM-Alex	IDE_{det}+CWS [Zhe16b]	20.0	48.2	78.8	20.5	48.3	78.8	20.5	48.3	78.8			
IAN (Resnet-101) [Xia19]		23.0	61.9										
DPM	PSE	29.3	65.1	88.3	31.7	65.0	88.2	32.4	64.5	87.5	32.6	63.9	87.0
DPM	PSAE	**31.2**	**66.4**	**89.1**	**34.1**	**66.3**	**89.0**	**34.8**	**65.9**	**88.6**	**34.8**	**65.7**	**87.9**

Table 5.11: Comparison of inference speed (in ms) for multiple CNN backbone architectures in dependence of batchsize. The Inception* architecture corresponds to the model used in Chapter 6.

Architecture	1	2	4	8	16
Inception*	4.54	2.44	1.65	1.06	0.86
Inception-V4	18.96	10.61	6.53	4.85	4.10
ResNet-50	5.10	2.87	1.99	1.65	1.49
ResNet-100	8.90	5.16	3.32	2.69	2.42

specifically designed for fast inference, an analysis of their runtime is provided to better judge constraints in real-world settings.

As depicted in Figure 3.1, a re-id approach consists of an offline and an online component, as well as possible person detection and tracking stages earlier in the pipeline. For the purposes of this analysis, detections in the form of pose information are presumed to be available. During the offline stage, features must be computed for all person images or tracks in a database. This process can happen prior to any actual re-id query and is thus less time-sensitive. However, real-time processing of the video streams originating from a camera network is often desirable. Table 5.11 shows average inference times of the proposed combined re-id model per person image. Experiments were carried out on an NVIDIA Titan X(P) GPU and inference speed was averaged over 100 forward passes after a warmup period of 20 passes.

While the runtime of the offline stage of a re-id system depends largely on the network architecture and amount of network parameters, the online stage is most strongly impacted by the dimensionality of the feature embedding. Table 5.12 gives an overview of matching speeds, i.e. distance computation, of a query to different database sizes and different feature dimensions. Experiments were carried out on a server with a Xeon E5-2650 2.20GHz CPU and 256GB of memory and averaged across 50 runs. Only a single core was used for distance computation. The online stage further requires a single forward-pass to compute the feature embedding of the query image but for realistically sized databases, this is often a negligible aspect.

Table 5.12: Comparison of runtime (in ms) for matching a query feature against a database of varying size and different embedding dimensions.

Feature Dim.	1k	10k	100k	1M	10M
128	0.2	1.2	39.4	393.7	3.98s
256	0.3	3.4	64.6	643.3	6.49s
1024	0.9	25.9	256.3	2.11s	18.09s

The results show that online matching speed is not a major factor. Single core speeds are already fast and distance computation is trivially parallelized across multiple cores or machines. In a real-world system, data access time will likely be the dominating factor, not matching speed. Inference time during the offline stage is a more important factor. For example, not taking into account previous detection or tracking stages, processing 1 million person detections during the offline stage with even the fastest reported network architecture would require approximately 17 minutes, while the resulting feature database could be searched in at most a few seconds.

Chapter 6

Domain Prototype Learning

Many re-id approaches, particularly those which automatically learn feature representations from data, are well adapted to the characteristics which are included in their training data. However, data with previously unseen or underrepresented characteristics can result in arbitrary outcomes. In practice, this situation can occur frequently. Examples for this are the deployment of re-id methods to new scenes or simply changes within existing camera systems, such as addition of new cameras, repositioning, or updates to newer models.

A popular strategy to address these concerns is the application of unsupervised learning [Rad16, Pen16]. Here, unannotated data from the target or changed scenario is used to adapt an existing model to new characteristics. However, model adaptation can take time, may require specialized or more powerful hardware, and may also require expert guidance for parameter settings. These requirements can often not be fulfilled in day-to-day use of a practical system. Particularly, if characteristics of the camera network data change frequently, this becomes a costly task with reduced practical usefulness.

In this chapter a two-stage framework for re-id is proposed to automatically discover visual domains in large amounts of diverse data and use them to learn

feature embeddings for re-id. This here described framework was previously published in [Sch17a].

- In the first stage, data from many diverse re-id datasets is pooled to capture a large degree of visual variation. Then, clustering based on feature learning in CNNs is applied to automatically discover dominant sets of visual characteristics, termed *prototype domains*.
- In the second stage of the approach, CNNs are used to learn feature embeddings for each of the prototype domains. A separate embedding is learned for each domain. At test time, a *domain-sensitive selection* process matches the query image to its closest visual domain and chooses the feature embedding learned on that domain to perform re-id.

The motivation behind this idea is to allow the domain embeddings to focus on specific details, which are important for this individual prototype domain, while ignoring those of other domains. For example, an embedding learned for a domain, which predominantly contains people with dark clothes, does not need to encode information relevant to distinguishing a person dressed in light blue colors from a person dressed in light green clothes. By focusing on a specialized domain, the domain perceptive embedding can then focus on learning more subtle discriminative characteristics among similar visual appearances.

6.1 Divergent Data Sampling

A key requirement for a meaningful domain discovery is *divergent data sampling* which aims to provide a large range of realistic visual variation. In order to achieve such a high degree of variation, a number of publicly available re-id datasets are pooled into a new, large dataset for domain discovery. Ten datasets are combined, which together contain images of 4,786 different persons with a total of 41,380 bounding boxes from 54 different camera views. Of these bounding boxes 27,283 were manually annotated and 14,097 are obtained by a person detector. Table 6.1 shows the sources used to construct

the proposed domain discovery dataset. All bounding boxes are resized to a uniform size of 160×80 pixels. This divergent data sampling provides a good basis to discover characteristic domains, which cover a large and diverse spectrum of possible variation in visual appearance of persons.

Table 6.1: Ten sources for the re-id domain discovery dataset. The data consists of manually labeled bounding boxes (M-BBoxes) as well as automatic person detections (A-BBoxes).

	Persons	Cameras	M-BBoxes	A-BBoxes
HDA [Fig14]	85	13	850	-
GRID [Loy10]	250	8	500	-
3DPeS [Bal11a]	200	8	1,012	-
CAVIAR4REID [Che11]	72	2	1,221	-
i-LIDS [Bra06]	119	2	476	-
PRID [Hir11]	200	2	400	-
VIPeR [Gra08]	632	2	1,264	-
SARC3D [Bal11b]	50	1	200	-
CUHK2 [Li13]	1,816	10	7,264	-
CUHK3 [Li14]	1,360	6	14,096	14,097
Total	4,786	54	27,283	14,097

6.2 Prototype Domain Discovery

Deep learning based clustering is used to discover prototype domains from the multi-source pooled dataset. The method is based on the concept of unsupervised deep embedding space learning proposed in [Xie16] and adapted to utilise the available person ID labels from the re-id datasets. The *self-supervised* deep learning clustering model alternates between two steps:

1. Applying conventional clustering, such as k-means, within the feature space of a CNN embedding to identify clusters.

2. Training of the underlying CNN to further adapt the feature embedding based on the cluster assignment.

These two iteratively performed steps enable the model to discover meaningful partitions in the data by not just clustering but also refining the underlying feature space simultaneously.

6.2.1 Initialization

In the absence of annotations, prior to the iterative clustering approach, the CNN could be initialized through training as an auto-encoder (AE), as is proposed in [Xie16]. However, this is likely to result in the domain discovery process focusing on global visual differences between the various datasets, rather than more subtle cues directly relating to the visual appearance of the depicted persons. Thus, an alternative initialization method is proposed. The model is trained for re-id using all of the person ID labels available in the data. The last, fully connected layer of the network is set to 4,786 dimensions and the network is trained using person ID labels in a one-hot encoding and a softmax cross-entropy loss for person ID classification. Two options are considered for the architecture of the model:

- The ResNet-50 architecure which has proven to provide a very strong baseline for re-id.
- A proposed custom Inception-based architecture, which contains fewer parameter than the ResNet-50 model, can be trained more quickly, and has a lower dimensional embedding. More details of this architecture are given in Table 6.2 and Section 6.3.1.

The multi-source dataset ensures that the influence of any particular dataset's bias on the initial feature embedding before clustering is reduced. Moreover, data augmentation is applied by cropping and flipping the images, and also ensuring unbiased sampling by selecting images from different original datasets with equal frequency. Image cropping is performed by resizing an image to 30×10 pixels larger than the network requires and randomly cropping the

image down to the expected size. Through data augmentation, the hypothetical data pool size is increased by a factor of 600. Uniform sampling from each of the original datasets results in more data augmentation on the smaller data sources, in order to prevent the resulting domains from being dominated by the larger datasets.

6.2.2 Iterative Domain Discovery

After this initial training, meaningful weights, which extract person information across the pooled dataset, have been obtained. For the domain discovery process, the person ID softmax cross-entropy loss layer of the CNN is then replaced by a softmax cross-entropy loss, which corresponds to the desired number of clusters or prototype domains. As the number of domains will have a direct impact on training times and offline processing speed, small numbers of clusters are preferable. After a supervised initialization, the domain discovery now continues in a self-supervised manner. An initial clustering is performed based on the learned re-id feature representation. The assigned cluster IDs are then used as image labels for the softmax cross-entropy loss. Further training of the CNN adapts the underlying feature space to better separate the clusters. The iterative process terminates when only few person images change their cluster assignment and stable domain borders have been established. This joint process of cluster assignment and learning of the underlying representation space gives the model a high flexibility in identifying meaningful subsets in the pooled data.

The initialization of weights through training of a re-id network is crucial to the success of the prototype domain discovery. The re-id training ensures that the initial model does *not* react strongly to the dataset biases present in the feature pool. This prevents the clustering from simply discovering trivial dataset boundaries as prototype domain boundaries and instead lets the model focus more on the content of each person bounding box. This is illustrated in Figure 6.2, which shows different learned prototype domains (i.e., clusters in the embedding feature space) with their corresponding images for initialization by auto-encoder or the proposed re-id initialization.

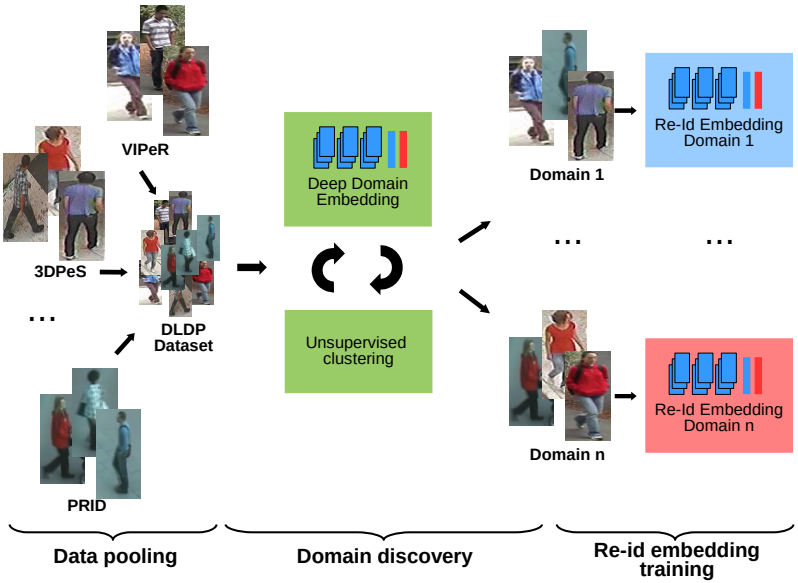

Figure 6.1: Overview of the proposed domain discovery process. An iterative process of clustering and CNN feature learning guides the discovery of meaningful subsets in the pooled data. Re-id models specific to the data characteristics in these subsets can then be trained.

6.2.3 Training Details

For training of the deep clustering model a low initial learning rate in the order of 10^{-4} is important. This ensures that the cluster embedding does not deviate too quickly from its re-id label sensitive initialization. Given the initial embedding, 25 runs of k-means clustering are performed in the embedding space. The result with the best separated and compact clusters is selected for the next refinement of the embedding. This ensures stability of the iterative training process. The refinement (fine-tuning) of the embedding CNN is then performed for a further 10,000 training iterations. The learning rate of the

Figure 6.2: Example domains discovered by the approach using the proposed initialization with a re-id net (top 3 rows, supervised initialization) and initialization by weights learned through autoencoding (AE) (bottom 3 rows, unsupervised initialization). The re-id initialization leads to more semantically meaningful domains (e.g., light-colored, yellow and blue clothing). The AE initialization is strongly influenced by dataset bias and learns domains corresponding to datasets (e.g., CAVIAR4REID, 3DPeS, PRID).

embedding is lowered by a factor of 0.1 every two iterations of the discovery process. This iterative process is repeated until less than 1% of images change their cluster assignments.

6.3 Domain Perceptive Re-Id Models

The second stage of the proposed re-id framework consists of training a domain-sensitive re-id model for each prototype domain. For this, one feature embedding for each of the discovered clusters is to be trained with all person IDs present in the data subset. First, a common generic baseline re-id model is trained on all available data without consideration of the domains. The individual domain models are then created by fine-tuning this baseline model.

6.3.1 Baseline Re-Id Model

Rather than relying on established large ImageNet-pretrained models, a more lightweight architecture is proposed, which is trained from scratch. The architecture is detailed in Table 6.2 and trained on all available training data to learn a generic feature embedding without domain specific adaptation. The architecture is inspired by that described in [Xia16a]. However, the model was improved by adding a fourth convolutional layer to enhance the feature representation at the beginning of the network and an increase in the final feature dimension results in more nuanced features, which lead to improved accuracy. Due to its reduced size, the baseline model is trained for just 60,000 iterations. The initial learning rate is set to 0.1 and divided by 10 after every 20,000 iterations. The 512 dimensional layer (fc feat in Table 6.2) just before the loss is used as feature embedding for person re-id. The resulting features are compared using cosine distance. The same baseline model is also used as initialization for the domain discovery approach described in Section 6.2.

6.3.2 Domain Embeddings Training

In order to learn feature embeddings focused on each of the domains, suitable domain-specific training data needs to be created. For any person ID in a given domain all of that person's images are selected and added to the training data for the domain. This happens regardless of whether all images

Table 6.2: The lightweight model architecture for prototype domain discovery and baseline re-id model.

name	patch size, stride	output dim	# filters
input		$3 \times 160 \times 64$	
conv 1-4	3×3, 1	$32 \times 160 \times 64$	
pool	2×2, 2	$32 \times 80 \times 32$	
inception 1a		$256 \times 80 \times 32$	64
inception 1b	stride 2	$384 \times 40 \times 16$	64
inception 2a		$512 \times 40 \times 16$	128
inception 2b	stride 2	$768 \times 20 \times 8$	128
inception 3a		$1024\times 20 \times 8$	128
inception 3b	stride 2	$1536\times 10 \times 4$	128
fc feat		512	
fc loss		#person ids	

of that person were originally assigned to this domain by the domain discovery process. Thus, person images may be assigned to several domains and the training data of different domains can partially overlap. The addition of all person images for a given ID is necessary, in order to provide the CNNs with sufficient visual variation within each person class. The partial overlap between training sets is not a drawback and may even help compensate error cases, when the wrong embedding is chosen for a new query. This data sampling method allows the domain models to specialize and focus particularly on the visual cues relevant to persons from their domain while not having to also learn how to distinguish persons from different domains.

Starting from the re-id baseline model, the networks are trained individually for each domain relying only on the corresponding data pool. The dimension of the final softmax layers is adapted accordingly. For each domain training continues for 30,000 iterations at an initial learning rate of 10^{-3}. The input images are resized to a size of 210×70 pixels. Data augmentation is then

performed by randomly flipping images and randomly cropping them to a final input size of 180×60 pixels. Similar to [Wu16b] hard negative mining is applied by selecting misclassified training images and fine-tuning each net on these difficult cases for a further 10,000 iterations at a reduced learning rate of 10^{-5}.

6.3.3 Automatic Domain Selection

The combination of domain classifier model and domain specific re-id models allows for a flexible query-adaptive deployment strategy. During model deployment, a probe person image is first matched to its closest domain. This can be done using the deep clustering model (Section 6.2) which generates probability values for cluster assignment. The corresponding domain's re-id model is then used to rank the gallery images by computing the 512 dimensional embedding and using cosine distance for matching the query image. An overview of this query-based selection process is given in Figure 6.3. This model selection process achieves an adaptation of the framework at query level. It requires only a single additional inference pass of the clustering CNN model and can thus be performed without significant delay for the operator.

Note that this approach is purely *inductive* when applied in a cross-domain setting. It does not require any training data (labelled or unlabelled) from the target domain, and is yet able to adapt based on the presented query images. The described method thus requires less data than unsupervised learning approaches and can be integrated into a practical system without requiring specialized knowledge from operating personnel or specialized hardware for further training. The approach is particularly well suited to scenarios in which no fixed set of camera views is available (i.e. no fixed domain borders are specified).

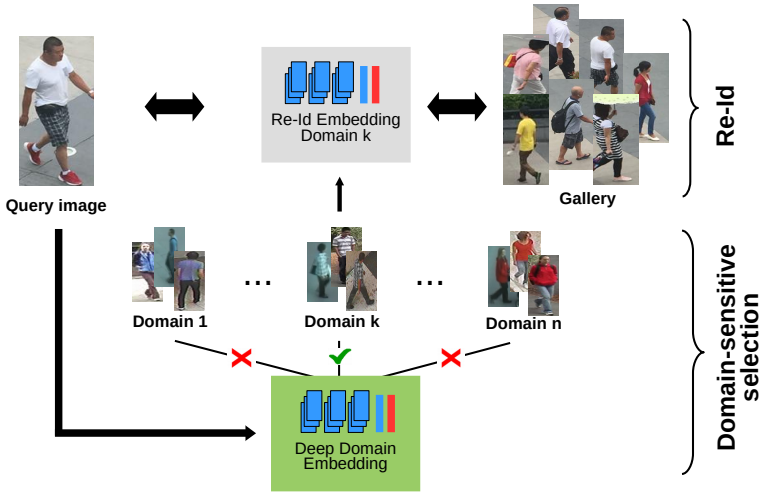

Figure 6.3: Test-time use of the proposed re-id framework. The query image is matched to the most closely related prototype domain by use of the clustering CNN. After this, the specialized re-id model of that domain is applied to perform the ranking task.

6.4 Evaluation

The proposed model was developed with practical application in mind. Thus, for evaluation two datasets are chosen, which allow for a range of different evaluation settings that closely mirror a real-world re-id scenario.

Datasets: The CUHK-SYSU [Xia16b] and PRW [Zhe16b] are both independent/unseen from the ten multi-source data pool used to construct the domain discovery training dataset. Both datasets contain a large number of viewing

angles. CUHK-SYSU consists of pedestrian images collected by handheld cameras as well as scenes from movies and TV series. The PRW dataset was collected with six fixed cameras on a campus environment. The datasets contain 8,432 and 932 person IDs and 99,809 and 34,304 bounding boxes, respectively. Importantly, both datasets provide full camera images. Thus, a combination of person detector and re-id model will have to be applied and the re-id stage will be subject to occlusion, bounding box misalignment and large changes in person resolution. Example images of both datasets are depicted in Figure 6.4. The high diversity and challenging nature of the two datasets are a well suited measure for the generalization capability of the proposed approach, as well as its ability to handle large amounts of varying views, indoor and outdoor scenes, and possible detection errors.

Figure 6.4: Dataset used in the evaluation of the domain selection re-id model. The top row shows fixed camera images from the PRW dataset and the bottom row mobile camera images and TV scenes from the CUHK-SYSU dataset.

Evaluation protocol: A central objective of the proposed approach is *not* to require any training data on the target domain for the re-id task. To that end, in the experiments we only used the test part of both datasets. The CUHK-SYSU dataset contains a fixed set of 2,900 query persons and gallery sets of multiple sizes (at most 6,978 full images). The PRW dataset contains a fixed query set of 2,057 bounding boxes and a gallery size of 6,112 test images. In both datasets each gallery image contains multiple persons and an automatic person detector may generate additional false positive bounding boxes. The original evaluation protocols from [Xia16b] and [Zhe16b] are applied, respectively, and existing evaluation code is used where available. Both datasets contain many persons without an ID in the galleries, i.e., the re-id tasks in these datasets have a large but unspecified number of distractors to handle. mAP and rank-1 accuracy are used as evaluation metrics.

Domain analysis: In Table 6.3 the influence of the number of chosen domains on the accuracy of the domain selection approach is evaluated. The setting for a single domain (k=1) corresponds to a straightforward baseline model, which uses the entire pooled data for training of a re-id embedding. For few domains (k=2) the resulting re-id models perform less accurate than the baseline. This is likely due to the low degree of specialization in the domains, which leads to the resulting models merely being weaker versions of the baseline model. Given an increasing number of domains, the advantage of the domain selection becomes greater until it saturates around eight domains. The overall trend is similar for both backbone architectures. The ResNet-50 model is more strongly impacted by using two clusters, compared to the baseline. Accuracy only gradually improves and is not as high as that of the Inception model. The ResNet-50 has more than five times more parameters than the Inception model. A reason for its weaker performance may lie in the smaller amounts of training data available for each domain, which can lead to overfitting in large models. For further experiments, the proposed inception architecture and eight domains are chosen.

Table 6.4 shows how often the domain selection process succeeds in choosing the correct domain embedding. For this, each domain embedding was evaluated for each query of the CUHK-SYSU dataset at the gallery 100 setting. The

Table 6.3: Effect of prototype-domain numbers (k) on re-id rate, using CUHK-SYSU with the gallery 100 setting.

Architecture		k=1	k=2	k=4	k=6	k=8
Inception	mAP	68.4	67.1	71.4	72.6	74.0
	rank-1	70.3	68.7	73.3	75.1	76.7
ResNet-50	mAP	69.9	66.5	67.5	70.8	72.9
	rank-1	71.0	64.3	67.7	72.1	75.9

Table 6.4: Relative number of cases in which each of the domain embeddings is the optimal choice, compared to the number of cases in which it is automatically chosen. The experiment was performed on CUHK-SYSU with the gallery 100 setting

	1	2	3	4	5	6	7	8
Correct	0.12	0.43	0.01	0.00	0.05	0.31	0.04	0.04
Chosen	0.10	0.45	0.03	0.00	0.02	0.29	0.05	0.06

optimal embedding was determined and compared to the automatic choice made by the domain classifier. The table shows a high correlation between the classifier's choice and the optimal choice.

Table 6.5: DLDP re-id performance comparison against both supervised (KISSME, IDNet, Person Search) and unsupervised (Euclidean, BoW) methods on the CUHK-SYSU dataset.

		mAP	rank-1
GT	Euclidean [Xia16b]	41.1	45.9
	KISSME [Koe12]	56.2	61.9
	BoW [Zhe15b]	62.5	67.2
	IDNet [Xia16b]	66.5	71.1
	Baseline Model	68.4	70.3
	DLDP	**74.0**	**76.7**
Detections	Person Search [Xia16b]	55.7	62.7
	Person Search rerun	55.79	62.17
	DLDP (SSD VOC300)	49.53	57.48
	DLDP (SSD VOC500)	**57.76**	**64.59**
	DLDP ([Xia16b] detections)	**66.76**	**71.93**

Table 6.6: Domain-adaptive re-id performance on the PRW dataset in comparison to state-of-the-art. All results are obtained by considering 5 bounding boxes per image. Note that all approaches except ours were trained (supervised) on the PRW dataset.

		mAP	rank-1
DPM	IDE [Zhe16b]	13.7	38.0
	IDE_{det} [Zhe16b]	**18.8**	**47.7**
	BoW + XQDA [Zhe16b]	12.1	36.2
	Baseline Model	12.9	36.5
	DLDP	15.9	45.4
SSD	BoW + XQDA (SSD VOC300)	6.8	26.6
	DLDP (SSD VOC300)	10.1	35.3
	DLDP (SSD VOC500)	**11.8**	**37.8**

Comparison with the state-of-the-art: To demonstrate the effectiveness of the approach, it is compared directly to the state-of-the-art reported in [Xia16b] and [Zhe16b], using both manually labeled person bounding boxes (i.e. ground truth) and automatically detected bounding boxes. Since the approach relies on deep learning domain prototypes, it is abbreviated DLDP in the tables. Results on the CUHK-SYSU dataset for gallery sizes of 100 images are given in Table 6.5. The baseline re-id model given manually labeled person bounding boxes as input outperforms not only [Xia16b] using conventional image features but also the deep IDNet model, which has the advantage of being trained on the CUHK-SYSU dataset, at rank-1 by 1.9%. The reason is likely a combination of the deeper 10 layer network architecture, the use of inception layers and batch normalization. The domain adaptive model, given manually labeled person bounding boxes, outperforms [Xia16b] by 7.5% and 5.6% in mAP and rank-1, respectively, which is a further improvement of 6% in both mAP and rank-1 over the baseline model. This indicates that the domain selection model for prototype-domain adaptive re-id is more effective than a direct model transfer.

For automatic detection of person bounding boxes, the SSD VOC500 person detector [Liu16] is applied. For re-id given these automatic detections, the

domain-adaptive model outperforms the state-of-the-art person search deep model [Xia16b] by 2.06% and 1.89% on mAP and rank-1, respectively. This is achieved despite the critical difference that the person search CNN model [Xia16b] was trained jointly for person detection and re-id using part of the CUHK-SYSU dataset. In contrast, the proposed model does not benefit from training detectors in the target domain, nor fine-tuning the re-id model on the target domain. If the domain-adaptive model is applied to the detections of the person search model, performance is further increased to 66.76% mAP and 71.93% rank-1 accuracy.

For the evaluation on the PRW benchmark, the domain-adaptive approach is compared to a baseline using BoW features and XQDA metric learning [Lia15a] and two deep feature embeddings IDE and IDE$_{det}$ from [Zhe16b], which are based on the AlexNet [Kri12] architecture, trained on ImageNet and fine-tuned for re-id on PRW. For person detection, both the DPM person detector [Fel10] trained on the INRIA dataset [Dal05] provided by [Zhe16b] and the SSD detector are used for a fair comparison. The results are shown in Table 6.6. All reported results were obtained by considering five bounding boxes per gallery image, which is the value at which the methods reported in [Zhe16b] perform best. It is evident that the SSD detectors decrease re-id performance for all models. This is at first a surprising observation. However, the annotations of the PRW dataset were created semi-automatically with the help of the DPM detector. Thus, the DMP detector has an uncommonly high accuracy on this dataset. The domain-adaptive model outperforms both the BOW+XQDA baseline and the deep IDE feature embedding reported in [Zhe16b], when an identical DPM person detector was used, by 2.2% and 7.4% in mAP and rank-1, respectively. The improved deep IDE$_{det}$ embedding of [Zhe16b] is trained by fine-tuning the AlexNet first for person and background classification, followed by further fine-tuning for re-id. It outperforms the domain-adaptive approach by 2.9% and 2.3% in mAP and rank-1 accuracy, respectively. However, the performance of the adaptive model remains competitive while neither requiring target-domain training, nor person-background classification.

The CUHK-SYSU dataset provides several gallery sets of varying size. Thus, the influence of larger numbers of distractors on the re-id accuracy can be

Table 6.7: Comparison of the domain-adaptive to [Xia16b] for different gallery sizes on CUHK-SYSU. Results of [Xia16b] were obtained using the provided code.

		50	100	500	1000	2000	4000	all (6978)
Person Search [Xia16b]		58.72	55.79	47.38	43.41	39.14	35.70	32.69
DLDP (SSD VOC500)		**60.8**	**57.7**	**49.2**	**45.2**	**41.4**	**38.1**	**35.2**
Person Search [Xia16b]		64.83	62.17	53.76	49.86	45.21	41.86	38.69
DLDP (SSD VOC500)		**67.6**	**64.6**	**57.0**	**52.9**	**49.2**	**46.1**	**43.1**

Table 6.8: Comparisons on the CUHK-SYSU occlusion and low resolution tests.

	Deep+Kissme [Xia16b]		ACF+BOW [Xia16b]		Person Search [Xia16b]		DLDP (SSD VOC500)	
	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1
Whole	39.1	44.9	42.4	48.4	55.7	62.7	**57.7**	**64.6**
Occlusion	18.2	17.7	29.1	32.1	**39.7**	**43.3**	38.9	39.0
LowRes	43.8	42.8	44.9	53.8	35.7	42.1	**41.9**	**49.0**

evaluated. Table 6.7 shows results of the DLDP model in comparison to the end-to-end person search CNN [Xia16b]. Overall, the DLDP model consistently outperforms the end-to-end person search model by a constant value of 2% in mAP regardless of gallery size. The DPDP model performs 3% better in rank-1 accuracy for low gallery sizes of 50 images (corresponding to 256 bounding boxes) and this margin increases up to 5.4% in rank-1 for the largest possible gallery of all 6978 images (i.e. 36,984 bounding boxes). This suggests that the proposed model is less sensitive to increase in gallery size, even without benefiting from learning on the target data.

Lastly, the DLDP model is evaluated for the effects of occlusion and low-resolution probe images. The CUHK-SYSU dataset contains subsets for this purpose, which were created by sampling heavily occluded probe images and those 10% probe images with the lowest resolutions, respectively. The gallery size for this evaluation is fixed at 100 images. Table 6.8 gives results using the SSD VOC500 detector, which are again compared th the person search model. It can be noted that an occluded probe image causes more difficulty for re-id than that of low-resolution imagery. This result is expected, as the

complete occlusion of potentially discriminative information is harder to compensate for than the gradual overall degradation of details caused by lower image resolutions. At low-resolutions, the DLDP model suffers only a 15% loss in mAP and rank-1, compared to a 20% decrease of the end-to-end person search model. For occlusions, the reported results of the end-to-end person search model are reduced by 16.0% mAP and 19.4% rank-1 and thus less affected than the DLDP model whose performance is reduced by 18.8% in mAP and 25.6% in rank-1.

In summary, the domain-adaptive model performs significantly stronger than a direct model transfer and outperforms several state-of-the-art re-id methods on both the CUHK-SYSU and the PRW benchmark datasets. Even some methods, which benefit from supervised training on the target domain, are outperformed. In challenging situations, such as increasing gallery size, low resolutions, or occluded query images, the model performs close to or even better than existing models, which were adapted to the data's characteristics. Qualitative examples on the CUHK-SYSU and PRW datasets are shown in Figures 6.5 and 6.6, respectively. Note that the incorrect results for all queries have a color composition or clothing configuration that is reasonably similar to the query image. In particular, the approach understandably ranks near-identically looking people (PRW, row 2) very high. In the failure case on PRW the model appears to focus on the structural pattern created by the bikes in combination with white-dressed persons.

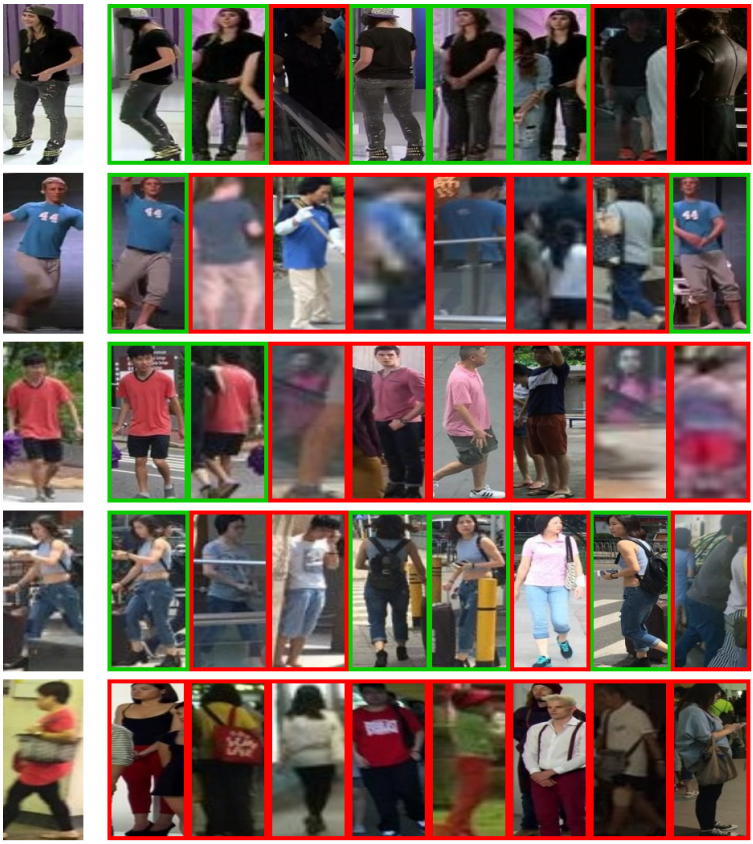

Figure 6.5: The top 8 re-id matches by the domain-adaptive model on the CUHK-SYSU test data for a set of five randomly chosen queries from the 100 image gallery setting. The bottom example shows failure cases when the model failed to find a match in the top 8 ranks.

Figure 6.6: Five randomly chosen queries on the PRW test data. Note, rank-2 and rank-3 in the "yellow T-shirt" example in are false matches even though they look very similar. The bottom example shows failure cases when the model failed to find a match in the top 8 ranks and likely focused on the structure of the bikes.

Chapter 7

Conclusion and Outlook

7.1 Conclusions

In this thesis two strategies for person re-identification using CNN models have been proposed. The first strategy relies on explicit modeling and inclusion of auxiliary information into the learning process of a re-id CNN. Semantic attributes and person pose information have been identified as promising candidates for this approach.

An attribute recognition CNN model was developed in order to provide the required attribute information. The experiments showed that a combination of local and global image information within the model can lead to notable improvements in the attribute recognition accuracy. Furthermore, pose information has proven to be a viable source as input for an attribute attention mechanism, which extracts the local information.

The attribute information was then used for re-id through training of a CNN-based feature representation, which learns complementary information to that contained in the attributes. This was achieved by a proposed modification of the triplet loss. The resulting feature combination showed a clear improvement in accuracy compared to conventionally trained CNN features

as well as common methods of fusing re-id features with attributes. However, the overall potential is limited by the low dimensionality of the attribute representation and the unavoidable degradation of accuracies though transfer to another dataset. Thus other elements of the attribute CNN were identified as more promising candidates for transfer.

For the use of pose information, the generated detections of body joint locations were encoded as additional input channels and provided to a re-id CNN. In addition to this, a classifier branch that determines coarse view information was added to the network in order to let later layers of the CNN specialize on the predicted view angles. The modeling of pose and view was shown to be complementary and was able to significantly increase re-id accuracy.

Both types of information were combined in a joint model and the results demonstrate that each element contributes to a further improved overall accuracy. The joint model also performs well on challenging evaluation settings, which mirror real-world requirements of modern re-id systems. This includes the ability to cope with errors from a previous person detection stage, scalability with increased numbers of persons in the gallery set, and video-based re-id.

The second strategy pursued in this work aimed at enabling a re-id system to adapt to new or changed target scenarios without the need for extensive and time consuming re-training. A re-id framework was proposed that consists of a set of CNN models. Each model is adapted to the specific characteristic of a prototype domain. These domains were automatically constructed from a large and diverse person dataset. Through selection of the closest matching model during test time, a degree of adaptation to each individual query was achieved. The resulting re-id accuracy was markedly improved compared to standard transfer of CNN models. Several existing baselines for unsupervised re-id approaches were outperformed and it could be shown that the correct specialized re-id model is chosen from the set with very high likelihood.

7.2 Outlook

The presented methods achieve a high degree of accuracy and address many of the core challenges involved with the development of a practical re-id system. However, several future research directions have been identified for further improvement of the proposed re-id methods.

- A more integrated *end-to-end model* could help combine and better adjust the different types of information used in this thesis. For example, the pose estimation stage could be trained as part of the re-id model. The gradient information from the re-id loss might then be able to influence the pose estimation segment in order to better avoid pose errors, which are damaging to the resulting re-id feature representation. In addition to exploiting such possible synergies during the learning process, an end-to-end model also has the potential to reduce computation time for the offline stage of the approach.
- Better use of *temporal information* is a promising direction, which is currently not addressed by the described models. While evaluations on video data were carried out, each image in a person track is treated separately and fusion across time is carried out by a simple averaging of the extracted features. A more explicit use of temporal information through, for example, recurrent networks can be expected to yield improved results. Particularly the auxiliary pose and attribute information could become a much more reliable support to the re-id model, when stabilized across the temporal context through pose-tracking or verification of temporal consistency of attributes.
- The *combination with face-based re-identification* as an additional source of auxiliary information may further improve accuracy and particularly robustness to changes in data characteristics. Face matching in the typical surveillance scenarios depicted in re-id datasets used to be unpromising, due to the low resolution, which

did not allow for face detection or matching. However, recently developed face detectors [Hu17] are able to detect even tiny faces with impressive accuracy. Furthermore, a first dataset for development of ultra low-resolution face matching methods has recently been published [Che18]. While the face-based representation can not be assumed to achieve such high confidences as are known from higher-resolution data, the added cue may be a valuable addition to re-id approaches. For example, confident individual face-based matches may be used to enhance the query set of a conventional re-id approach.

From a more global perspective, the accuracy of re-id approaches has improved significantly over recent years. Many current models achieve rank-1 accuracies upwards of 90% [Li18b, Zha17a], which can in some cases be considered better than human accuracy [Zha17a]. The same trend could be observed in the field of face recognition where many models now outperform human accuracy [Lu17a, Sch15b]. Consequently, future re-id research will likely begin to focus on more specialized tasks.

- Most current re-id datasets and evaluation protocols implicitly assume that for each query there is at least one correct match in the gallery. In contrast to this closed world assumption, *open world* re-id is a much more challenging problem. Here, re-id systems need to either find correct matches to a query image or decide, if no correct match is present. This is a much more difficult decision to make and accuracies of the few existing approaches [Lia14, Zhe16c, Zhu18] are consequently much lower.
- Re-id of persons across *larger domain gaps* will be an interesting direction of future research. An example use case for this is the matching of people between visual and infrared cameras. Night time outdoor surveillance systems often rely on infrared imagery and an automated system for matching between day and night or indoor and outdoor areas will have to bridge this gap. Some first datasets [Wu17, Ngu17] and approaches [Ye18, Dai18] for this specialized re-id use case already exist.

- The rising accuracy in attribute recognition in combination with more detailed sets of attributes allow for a purely text-based re-id query. This problem may occur when only a witness description of a person of interest is available. Such *cross-modal text-to-image matching* has already been applied for fashion applications where datasets [Liu12] are of higher resolution and more nuanced description of clothing attributes exist [Lae17, Li17c]. However, first works aim at employing similar methods for the task of person re-identification [Yin17].

Bibliography

[Abd15] Abdulnabi, Abrar H; Wang, Gang; Lu, Jiwen and Jia, Kui: "Multi-task CNN model for attribute prediction". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1949–1959 (cit. on p. 31).

[Ahm15] Ahmed, Ejaz; Jones, Michael and Marks, Tim K: "An improved deep learning architecture for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3908–3916 (cit. on p. 15).

[Bai16] Bai, Song and Bai, Xiang: "Sparse contextual activation for efficient visual re-ranking". In: *IEEE Transactions on Image Processing* 25.3 (2016), pp. 1056–1069 (cit. on p. 23).

[Bal11a] Baltieri, Davide; Vezzani, Roberto and Cucchiara, Rita: "3dpes: 3d people dataset for surveillance and forensics". In: *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM. 2011, pp. 59–64 (cit. on p. 103).

[Bal11b] Baltieri, Davide; Vezzani, Roberto and Cucchiara, Rita: "Sarc3d: a new 3d body model for people tracking and re-identification". In: *International Conference on Image Analysis and Processing*. 2011 (cit. on p. 103).

[Bäu13] Bäuml, Martin; Tapaswi, Makarand and Stiefelhagen, Rainer: "Semi-supervised Learning with Constraints for Person Identification in Multimedia Data". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013 (cit. on p. 2).

[Bou11] Bourdev, Lubomir; Maji, Subhransu and Malik, Jitendra: "Describing people: A poselet-based approach to attribute classification". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 1543–1550 (cit. on pp. 30, 37).

[Bra06] Branch, HOSD: "Imagery library for intelligent detection systems (i-lids)". In: *The Institution of Engineering and Technology Conference on Crime and Security*. 2006, pp. 445–448 (cit. on p. 103).

[Cao17] Cao, Zhe; Simon, Tomas; Wei, Shih-En and Sheikh, Yaser: "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CVPR*. 2017 (cit. on p. 52).

[Che11] Cheng, Dong Seon; Cristani, Marco; Stoppa, Michele; Bazzani, Loris and Murino, Vittorio: "Custom pictorial structures for re-identification." In: *Bmvc*. Vol. 1. 2. Citeseer. 2011, p. 6 (cit. on pp. 21, 29, 103).

[Che12] Chen, Huizhong; Gallagher, Andrew and Girod, Bernd: "Describing clothing by semantic attributes". In: *European conference on computer vision*. Springer. 2012, pp. 609–623 (cit. on pp. 30, 37).

[Che16] Cheng, De; Gong, Yihong; Zhou, Sanping; Wang, Jinjun and Zheng, Nanning: "Person re-identification by multi-channel parts-based cnn with improved triplet loss function". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1335–1344 (cit. on p. 19).

[Che17] Chen, Yanbei; Zhu, Xiatian and Gong, Shaogang: "Person re-identification by deep learning multi-scale representations". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2590–2600 (cit. on p. 17).

[Che18] Cheng, Zhiyi; Zhu, Xiatian and Gong, Shaogang: "Surveillance Face Recognition Challenge". In: *arXiv preprint arXiv:1804.09691* (2018) (cit. on p. 124).

[Cho16] Cho, Yeong-Jun and Yoon, Kuk-Jin: "Improving person re-identification via pose-aware multi-shot matching". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1354–1362 (cit. on p. 21).

[Dai18] Dai, Pingyang; Ji, Rongrong; Wang, Haibin; Wu, Qiong and Huang, Yuyu: "Cross-Modality Person Re-Identification with Generative Adversarial Training." In: *IJCAI*. 2018, pp. 677–683 (cit. on p. 124).

[Dal05] Dalal, Navneet and Triggs, Bill: "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 886–893 (cit. on p. 116).

[Dav07] Davis, Jason V; Kulis, Brian; Jain, Prateek; Sra, Suvrit and Dhillon, Inderjit S: "Information-theoretic metric learning". In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 209–216 (cit. on p. 13).

[Den09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai and Fei-Fei, Li: "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee. 2009, pp. 248–255 (cit. on pp. 39, 45).

[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*. ACM. 2014, pp. 789–792 (cit. on p. 37).

[Den15] Deng, Yubin; Luo, Ping; Loy, Chen Change and Tang, Xiaoou: "Learning to recognize pedestrian attribute". In: *arXiv preprint arXiv:1501.00901* (2015) (cit. on p. 30).

[Den18] Deng, Weijian; Zheng, Liang; Ye, Qixiang; Kang, Guoliang; Yang, Yi and Jiao, Jianbin: "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 994–1003 (cit. on p. 20).

[Dib16] Diba, Ali; Mohammad Pazandeh, Ali; Pirsiavash, Hamed and Van Gool, Luc: "Deepcamp: Deep convolutional action & attribute mid-level patterns". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3557–3565 (cit. on p. 31).

[Din15] Ding, Shengyong; Lin, Liang; Wang, Guangrun and Chao, Hongyang: "Deep feature learning with relative distance comparison for person re-identification". In: *Pattern Recognition* 48.10 (2015), pp. 2993–3003 (cit. on p. 18).

[Don17] Dong, Qi; Gong, Shaogang and Zhu, Xiatian: "Multi-task curriculum transfer deep learning of clothing attributes". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2017, pp. 520–529 (cit. on p. 33).

[Dud12] Duda, Richard O; Hart, Peter E and Stork, David G: Pattern classification. John Wiley & Sons, 2012 (cit. on p. 43).

[Fab17] Fabbri, Matteo; Calderara, Simone and Cucchiara, Rita: "Generative adversarial models for people attribute recognition in surveillance". In: *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2017, pp. 1–6 (cit. on p. 33).

[Far10] Farenzena, Michela; Bazzani, Loris; Perina, Alessandro; Murino, Vittorio and Cristani, Marco: "Person re-identification by symmetry-driven accumulation of local features". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 2360–2367 (cit. on pp. 12, 21).

[Fel10] Felzenszwalb, Pedro F; Girshick, Ross B; McAllester, David and Ramanan, Deva: "Object detection with discriminatively trained part-based models". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2010), pp. 1627–1645 (cit. on pp. 29, 116).

[Fig14] Figueira, Dario; Taiana, Matteo; Nambiar, Athira; Nascimento, Jacinto and Bernardino, Alexandre: "The hda+ data set for research on fully automated re-identification systems". In: *Proceedings of the European Conference on Computer Vision*. 2014 (cit. on p. 103).

[Gar15] Garcia, Jorge; Martinel, Niki; Micheloni, Christian and Gardel, Alfredo: "Person re-identification ranking optimisation by discriminant context information analysis". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1305–1313 (cit. on p. 22).

[Ghe06] Gheissari, Niloofar; Sebastian, Thomas B and Hartley, Richard: "Person reidentification using spatiotemporal appearance". In: *null*. IEEE. 2006, pp. 1528–1535 (cit. on p. 12).

[Gki15] Gkioxari, Georgia; Girshick, Ross and Malik, Jitendra: "Actions and attributes from wholes and parts". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2470–2478 (cit. on p. 31).

[Gra08] Gray, Douglas and Tao, Hai: "Viewpoint invariant pedestrian recognition with an ensemble of localized features". In: *European conference on computer vision*. Springer. 2008, pp. 262–275 (cit. on pp. 6, 12, 29, 103).

[Guo17] Guo, Hao; Fan, Xiaochuan and Wang, Song: "Human attribute recognition by refining attention heat map". In: *Pattern Recognition Letters* 94 (2017), pp. 38–45 (cit. on p. 32).

[Hal15] Hall, David and Perona, Pietro: "Fine-grained classification of pedestrians in video: Benchmark and state of the art". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5482–5491 (cit. on p. 37).

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing and Sun, Jian: "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 46, 48).

[Her17] Hermans, Alexander; Beyer, Lucas and Leibe, Bastian: "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017) (cit. on pp. 19, 95).

[Hir11] Hirzer, Martin; Beleznai, Csaba; Roth, Peter M and Bischof, Horst: "Person re-identification by descriptive and discriminative classification". In: *Scandinavian conference on Image analysis*. Springer. 2011, pp. 91–102 (cit. on pp. 29, 103).

[Hir12] Hirzer, Martin; Roth, Peter M; Köstinger, Martin and Bischof, Horst: "Relaxed pairwise learned metric for person re-identification". In: *European Conference on Computer Vision*. Springer. 2012, pp. 780–793 (cit. on p. 14).

[Hu17] Hu, Peiyun and Ramanan, Deva: "Finding tiny faces". In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE. 2017, pp. 1522–1530 (cit. on p. 124).

[Hua17] Huang, Gao; Liu, Zhuang; Maaten, Laurens van der and Weinberger, Kilian Q: "Densely connected convolutional networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017 (cit. on pp. 46, 48).

[Ins16] Insafutdinov, Eldar; Pishchulin, Leonid; Andres, Bjoern; Andriluka, Mykhaylo and Schiele, Bernt: "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: *European Conference on Computer Vision*. Springer. 2016, pp. 34–50 (cit. on pp. 52, 60, 82).

[Iof15] Ioffe, Sergey and Szegedy, Christian: "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015) (cit. on p. 45).

[Jeg07] Jegou, Herve; Harzallah, Hedi and Schmid, Cordelia: "A contextual dissimilarity measure for accurate and efficient image search". In: *Computer Vision and Pattern Recognition,CVPR*. IEEE. 2007, pp. 1–8 (cit. on p. 23).

[Kan77] Kanade, Takeo: Computer Recognition of Human Faces, volume 47 of Interdisciplinary Systems Research. 1977 (cit. on p. 2).

[Kha14] Khamis, Sameh; Kuo, Cheng-Hao; Singh, Vivek K; Shet, Vinay D and Davis, Larry S: "Joint learning for attribute-consistent person re-identification". In: *European Conference on Computer Vision*. Springer. 2014, pp. 134–146 (cit. on p. 20).

[Kin14] Kingma, Diederik P and Ba, Jimmy: "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 44).

[Koe12] Koestinger, Martin; Hirzer, Martin; Wohlhart, Paul; Roth, Peter M and Bischof, Horst: "Large scale metric learning from equivalence constraints". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2288–2295 (cit. on pp. 13, 74, 114).

[Kri12] Krizhevsky, Alex; Sutskever, Ilya and Hinton, Geoffrey E: "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105 (cit. on pp. 39, 40, 45, 46, 116).

[Lae17] Laenen, Katrien; Zoghbi, Susana and Moens, Marie-Francine: "Cross-modal search for fashion attributes". In: *Proceedings of the KDD 2017 Workshop on Machine Learning Meets Fashion*. ACM. 2017 (cit. on p. 125).

[Lay12a] Layne, Ryan; Hospedales, Timothy M and Gong, Shaogang: "Person Re-identification by Attributes". In: *British Machine Vision Conference (BMVC)*. 2012 (cit. on pp. 20, 30).

[Lay12b] Layne, Ryan; Hospedales, Timothy M and Gong, Shaogang: "Towards person identification and re-identification with attributes". In: *European Conference on Computer Vision*. Springer. 2012, pp. 402–412 (cit. on p. 20).

[Lay14] Layne, Ryan; Hospedales, Timothy M and Gong, Shaogang: "Attributes-based re-identification". In: *Person Re-Identification*. Springer, 2014, pp. 93–117 (cit. on p. 20).

[Len15] Leng, Qingming; Hu, Ruimin; Liang, Chao; Wang, Yimin and Chen, Jun: "Person re-identification with content and context re-ranking". In: *Multimedia Tools and Applications* 74.17 (2015), pp. 6989–7014 (cit. on p. 22).

[Li12a] Li, Wei; Wu, Yang; Mukunoki, Masayuki and Minoh, Michihiko: "Common-near-neighbor analysis for person re-identification". In: *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE. 2012, pp. 1621–1624 (cit. on p. 22).

[Li12b] Li, Wei; Zhao, Rui and Wang, Xiaogang: "Human reidentification with transferred metric learning". In: *Asian Conference on Computer Vision*. Springer. 2012, pp. 31–44 (cit. on p. 29).

[Li13] Li, Wei and Wang, Xiaogang: "Locally aligned feature transforms across views". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3594–3601 (cit. on pp. 29, 103).

[Li14] Li, Wei; Zhao, Rui; Xiao, Tong and Wang, Xiaogang: "Deepreid: Deep filter pairing neural network for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 152–159 (cit. on pp. 15, 29, 103).

[Li15] Li, Dangwei; Chen, Xiaotang and Huang, Kaiqi: "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE. 2015, pp. 111–115 (cit. on pp. 31, 59, 64, 65).

[Li16a] Li, Dangwei; Zhang, Zhang; Chen, Xiaotang; Ling, Haibin and Huang, Kaiqi: "A richly annotated dataset for pedestrian attribute recognition". In: *arXiv preprint arXiv:1603.07054* (2016) (cit. on pp. 37, 38, 81, 83).

[Li16b] Li, Yining; Huang, Chen; Loy, Chen Change and Tang, Xiaoou: "Human attribute recognition by deep hierarchical contexts". In: *European Conference on Computer Vision*. Springer. 2016, pp. 684–700 (cit. on pp. 32, 37).

[Li17a] Li, Dangwei; Chen, Xiaotang; Zhang, Zhang and Huang, Kaiqi: "Learning Deep Context-Aware Features Over Body and Latent Parts for Person Re-Identification". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (cit. on p. 94).

[Li17b] Li, Dangwei; Chen, Xiaotang; Zhang, Zhang and Huang, Kaiqi: "Learning deep context-aware features over body and latent parts for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 384–393 (cit. on p. 22).

[Li17c] Li, Shuang; Xiao, Tong; Li, Hongsheng; Yang, Wei and Wang, Xiaogang: "Identity-aware textual-visual matching with latent co-attention". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE. 2017, pp. 1908–1917 (cit. on p. 125).

[Li17d] Li, Wei; Zhu, Xiatian and Gong, Shaogang: "Person Re-Identification by Deep Joint Learning of Multi-Loss Classification". In: *International Joint Conference of Artificial Intelligence*. 2017 (cit. on p. 94).

[Li17e] Li, Wei; Zhu, Xiatian and Gong, Shaogang: "Person re-identification by deep joint learning of multi-loss classification". In: *arXiv preprint arXiv:1705.04724* (2017) (cit. on p. 17).

[Li18a] 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)*. IEEE. 2018, pp. 1–6 (cit. on p. 31).

[Li18b] Li, Wei; Zhu, Xiatian and Gong, Shaogang: "Harmonious attention network for person re-identification". In: *CVPR*. Vol. 1. 2018, p. 2 (cit. on pp. 17, 94, 124).

[Li19] Li, Dangwei; Zhang, Zhang; Chen, Xiaotang and Huang, Kaiqi: "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios". In: *IEEE transactions on image processing* 28.4 (2019), pp. 1575–1590 (cit. on p. 64).

[Lia14] Liao, Shengcai; Mo, Zhipeng; Zhu, Jianqing; Hu, Yang and Li, Stan Z: "Open-set person re-identification". In: *arXiv preprint arXiv:1408.0872* (2014) (cit. on p. 124).

[Lia15a] Liao, Shengcai; Hu, Yang; Zhu, Xiangyu and Li, Stan Z: "Person re-identification by local maximal occurrence representation and metric learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2197–2206 (cit. on pp. 13, 14, 116).

[Lia15b] Liao, Shengcai and Li, Stan Z: "Efficient psd constrained asymmetric metric learning for person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3685–3693 (cit. on p. 14).

[Lin13] Lin, Min; Chen, Qiang and Yan, Shuicheng: "Network in network". In: *arXiv preprint arXiv:1312.4400* (2013) (cit. on p. 47).

[Lin17] Lin, Yutian; Zheng, Liang; Zheng, Zhedong; Wu, Yu and Yang, Yi: "Improving person re-identification by attribute and identity learning". In: *arXiv preprint arXiv:1703.07220* (2017) (cit. on pp. 21, 37, 95).

[Liu12] Liu, Chunxiao; Gong, Shaogang; Loy, Chen Change and Lin, Xinggang: "Person re-identification: What features are important?" In: *European Conference on Computer Vision*. Springer. 2012, pp. 391–401 (cit. on p. 125).

[Liu16] Liu, Wei; Anguelov, Dragomir; Erhan, Dumitru; Szegedy, Christian; Reed, Scott; Fu, Cheng-Yang and Berg, Alexander C: "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37 (cit. on p. 115).

[Liu17a] Liu, Hao; Feng, Jiashi; Qi, Meibin; Jiang, Jianguo and Yan, Shuicheng: "End-to-end comparative attention networks for person re-identification". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3492–3506 (cit. on p. 18).

[Liu17b] 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: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 350–359 (cit. on pp. 32, 37, 64).

[Liu17c] Liu, Yu; Yan, Junjie and Ouyang, Wanli: "Quality aware network for set to set recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5790–5799 (cit. on p. 94).

[Liu18a] Liu, Hao; Jie, Zequn; Jayashree, Karlekar; Qi, Meibin; Jiang, Jianguo; Yan, Shuicheng and Feng, Jiashi: "Video-based person re-identification with accumulative motion context". In: *IEEE transactions on circuits and systems for video technology* 28.10 (2018), pp. 2788–2802 (cit. on pp. 19, 20).

[Liu18b] Liu, Hao; Wu, Jingjing; Jiang, Jianguo; Qi, Meibin and Bo, Ren: "Sequence-based Person Attribute Recognition with Joint CTC-Attention Model". In: *arXiv preprint arXiv:1811.08115* (2018) (cit. on p. 33).

[Loy09] Loy, Chen Change; Xiang, Tao and Gong, Shaogang: "Multi-camera activity correlation analysis". In: *CVPR*. IEEE, 2009 (cit. on p. 29).

[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 (2010) (cit. on p. 103).

[Lu17a] Lu, Boyu; Zheng, Jingxiao; Chen, Jun-Cheng and Chellappa, Rama: "Pose-Robust Face Verification by Exploiting Competing Tasks". In: *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE. 2017, pp. 1124–1132 (cit. on p. 124).

[Lu17b] Lu, Yongxi; Kumar, Abhishek; Zhai, Shuangfei; Cheng, Yu; Javidi, Tara and Feris, Rogerio: "Fully-adaptive feature sharing in

multi-task networks with applications in person attribute classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5334–5343 (cit. on p. 33).

[Mat16] Matsukawa, ES Tetsu and Suzuki, Einoshin: "Person Re-Identification Using CNN Features Learned from Combination of Attributes". In: ICPR. 2016 (cit. on p. 21).

[McL16] McLaughlin, Niall; Martinez del Rincon, Jesus and Miller, Paul: "Recurrent convolutional network for video-based person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1325–1334 (cit. on pp. 19, 20).

[Mig12] Mignon, Alexis and Jurie, Frédéric: "Pcca: A new approach for distance learning from sparse pairwise constraints". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2666–2672 (cit. on p. 12).

[Mik99] Mika, Sebastian; Ratsch, Gunnar; Weston, Jason; Scholkopf, Bernhard and Mullers, Klaus-Robert: "Fisher discriminant analysis with kernels". In: *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop*. Ieee. 1999, pp. 41–48 (cit. on p. 14).

[Ngu17] Nguyen, Dat Tien; Hong, Hyung Gil; Kim, Ki Wan and Park, Kang Ryoung: "Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras". In: *Sensors* 17.3 (2017), p. 605 (cit. on p. 124).

[Ped13] Pedagadi, Sateesh; Orwell, James; Velastin, Sergio and Boghossian, Boghos: "Local fisher discriminant analysis for pedestrian re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3318–3325 (cit. on p. 12).

[Pen16] Peng, Peixi; Xiang, Tao; Wang, Yaowei; Pontil, Massimiliano; Gong, Shaogang; Huang, Tiejun and Tian, Yonghong: "Unsupervised cross-dataset transfer learning for person re-identification".

In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1306–1315 (cit. on p. 101).

[Qin11] Qin, Danfeng; Gammeter, Stephan; Bossard, Lukas; Quack, Till and Van Gool, Luc: "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 777–784 (cit. on p. 23).

[Rad16] Radenović, Filip; Tolias, Giorgos and Chum, Ondřej: "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples". In: *European conference on computer vision*. Springer. 2016, pp. 3–20 (cit. on pp. 18, 101).

[Rah17] Rahimpour, Alireza; Liu, Liu; Taalimi, Ali; Song, Yang and Qi, Hairong: "Person Re-identification Using Visual Attention". In: *arXiv preprint arXiv:1707.07336* (2017) (cit. on p. 22).

[Ris16] Ristani, Ergys; Solera, Francesco; Zou, Roger; Cucchiara, Rita and Tomasi, Carlo: "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking". In: *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*. 2016 (cit. on pp. 29, 73).

[Ros58] Rosenblatt, Frank: "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 39).

[Saq18] Saquib Sarfraz, M; Schumann, Arne; Eberle, Andreas and Stiefelhagen, Rainer: "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 420–429 (cit. on p. 94).

[Sar17a] Sarafianos, Nikolaos; Giannakopoulos, Theodore; Nikou, Christophoros and Kakadiaris, Ioannis A: "Curriculum learning for multi-task classification of visual attributes". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2608–2615 (cit. on p. 33).

[Sar17b] Sarfraz, M Saquib; Schumann, Arne; Wang, Yan and Stiefelhagen, Rainer: "Deep View-Sensitive Pedestrian Attribute Inference in an end-to-end Model". In: *British Machine Vision Conference (BMVC)*. 2017 (cit. on pp. 8, 60, 64, 65, 79).

[Sar18a] Sarafianos, Nikolaos; Xu, Xiang and Kakadiaris, Ioannis A: "Deep imbalanced attribute classification using visual attention aggregation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 680–697 (cit. on pp. 32, 64).

[Sar18b] Sarfraz, M Saquib; Schumann, Arne; Eberle, Andreas and Stiefelhagen, Rainer: "A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking". In: *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. 2018 (cit. on pp. 9, 23, 79).

[Sch15a] Schmidhuber, Jürgen: "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117 (cit. on p. 39).

[Sch15b] Schroff, Florian; Kalenichenko, Dmitry and Philbin, James: "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823 (cit. on pp. 18, 124).

[Sch15c] Schumann, Arne and Schuchert, Tobias: "Person Re-Identification in UAV Videos using Relevance Feedback". In: *Video Surveillance and Transportation Imaging Applications*. Vol. 9407. International Society for Optics and Photonics. 2015, 94070Z (cit. on p. 2).

[Sch17a] Schumann, Arne; Gong, Shaogang and Schuchert, Tobias: "Deep Learning Prototype Domains for Person Re-Identification". In: *Image Processing (ICIP), IEEE International Conference on*. IEEE. 2017, pp. 1767–1771 (cit. on pp. 9, 102).

[Sch17b] Schumann, Arne and Stiefelhagen, Rainer: "Person Re-Identification by Deep Learning Attribute-Complementary Information". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference on*. IEEE. 2017, pp. 1435–1443 (cit. on pp. 8, 69).

[Sch17c] Schumann, Arne and Stiefelhagen, Rainer: "Person Re-Identification by Deep Learning Attribute-Complementary Information". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE. 2017, pp. 1435–1443 (cit. on p. 94).

[Sha11] Sharma, Gaurav and Jurie, Frederic: "Learning discriminative spatial representation for image classification". In: *BMVC 2011-British Machine Vision Conference*. BMVA Press. 2011, pp. 1–11 (cit. on pp. 30, 37).

[She12] Shen, Xiaohui; Lin, Zhe; Brandt, Jonathan; Avidan, Shai and Wu, Ying: "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 3013–3020 (cit. on p. 22).

[She15] Shen, Yang; Lin, Weiyao; Yan, Junchi; Xu, Mingliang; Wu, Jianxin and Wang, Jingdong: "Person re-identification with correspondence structure learning". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3200–3208 (cit. on p. 13).

[Shi15] Shi, Zhiyuan; Hospedales, Timothy M and Xiang, Tao: "Transferring a semantic representation for person re-identification and search". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2015, pp. 4184–4193 (cit. on p. 20).

[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. 45, 46, 48).

[Son18] Song, Chunfeng; Huang, Yan; Ouyang, Wanli and Wang, Liang: "Mask-guided contrastive attention model for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1179–1188 (cit. on pp. 19, 94).

[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 p. 44).

[Su15] Su, Chi; Yang, Fan; Zhang, Shiliang; Tian, Qi; Davis, Larry S and Gao, Wen: "Multi-task learning with low rank attribute embedding for person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3739–3747 (cit. on p. 20).

[Su17a] Su, Chi; Li, Jianing; Zhang, Shiliang; Xing, Junliang; Gao, Wen and Tian, Qi: "Pose-driven Deep Convolutional Model for Person Re-identification". In: *Proceedings of the IEEE Conference on Computer Vision ICCV*. 2017, pp. 3960–3969 (cit. on pp. 17, 21, 94).

[Su17b] Su, Chi; Zhang, Shiliang; Yang, Fan; Zhang, Guangxiao; Tian, Qi; Gao, Wen and Davis, Larry S: "Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping". In: *Pattern Recognition* 66 (2017), pp. 4–15 (cit. on p. 20).

[Su18] Su, Chi; Zhang, Shiliang; Xing, Junliang; Gao, Wen and Tian, Qi: "Multi-type attributes driven multi-camera person re-identification". In: *Pattern Recognition* 75 (2018), pp. 77–89 (cit. on p. 21).

[Sud15] Sudowe, Patrick; Spitzer, Hannah and Leibe, Bastian: "Person attribute recognition with a jointly-trained holistic cnn model". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 87–95 (cit. on pp. 31, 37, 64, 65).

[Sud16] Sudowe, Patrick and Leibe, Bastian: "PatchIt: Self-Supervised Network Weight Initialization for Fine-grained Recognition." In: *BMVC*. Vol. 1. 2016, pp. 24–25 (cit. on p. 33).

[Sun17] Sun, Yifan; Zheng, Liang; Deng, Weijian and Wang, Shengjin: "SVDNet for Pedestrian Retrieval". In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 (cit. on p. 94).

[Sun18] Sun, Yifan; Zheng, Liang; Yang, Yi; Tian, Qi and Wang, Shengjin: "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 480–496 (cit. on p. 17).

[Sze15] Szegedy, Christian; Liu, Wei; Jia, Yangqing; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent and Rabinovich, Andrew: "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on pp. 47, 48, 64).

[Sze17] Szegedy, Christian; Ioffe, Sergey; Vanhoucke, Vincent and Alemi, Alexander A: "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." In: *AAAI*. 2017, pp. 4278–4284 (cit. on pp. 47, 48).

[Van09] Van De Weijer, Joost; Schmid, Cordelia; Verbeek, Jakob and Larlus, Diane: "Learning color names for real-world applications". In: *IEEE Transactions on Image Processing* 18.7 (2009), pp. 1512–1523 (cit. on p. 12).

[Var16a] Varior, Rahul Rama; Haloi, Mrinal and Wang, Gang: "Gated siamese convolutional neural network architecture for human re-identification". In: *European Conference on Computer Vision*. Springer. 2016, pp. 791–808 (cit. on p. 16).

[Var16b] Varior, Rahul Rama; Shuai, Bing; Lu, Jiwen; Xu, Dong and Wang, Gang: "A siamese long short-term memory architecture for human re-identification". In: *European Conference on Computer Vision*. Springer. 2016, pp. 135–153 (cit. on p. 20).

[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: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2285–2294 (cit. on p. 33).

[Wan17] Wang, Jingya; Zhu, Xiatian; Gong, Shaogang and Li, Wei: "Attribute recognition by joint recurrent learning of context and correlation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 531–540 (cit. on p. 33).

[Wan18a] Wang, Cheng; Zhang, Qian; Huang, Chang; Liu, Wenyu and Wang, Xinggang: "Mancs: A multi-task attentional network with curriculum sampling for person re-identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 365–381 (cit. on p. 94).

[Wan18b] Wang, Jingya; Zhu, Xiatian; Gong, Shaogang and Li, Wei: "Transferable joint attribute-identity deep learning for unsupervised person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2275–2284 (cit. on p. 21).

[Wei09] Weinberger, Kilian Q and Saul, Lawrence K: "Distance metric learning for large margin nearest neighbor classification". In: *Journal of Machine Learning Research* 10.Feb (2009), pp. 207–244 (cit. on pp. 13, 43).

[Wei17] Wei, Longhui; Zhang, Shiliang; Yao, Hantao; Gao, Wen and Tian, Qi: "Glad: Global-local-alignment descriptor for pedestrian retrieval". In: *Proceedings of the 25th ACM international conference on Multimedia*. ACM. 2017, pp. 420–428 (cit. on p. 22).

[Wei18a] Wei, Longhui; Zhang, Shiliang; Gao, Wen and Tian, Qi: "Person transfer gan to bridge domain gap for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 79–88 (cit. on p. 20).

[Wei18b] Wei, Longhui; Zhang, Shiliang; Gao, Wen and Tian, Qi: "Person Trasfer GAN to Bridge Domain Gap for Person Re-Identification". In: *Computer Vision and Pattern Recognition, IEEE Conference on*. 2018 (cit. on p. 29).

[Wu16a] Wu, Lin; Shen, Chunhua and Hengel, Anton van den: "Personnet: Person re-identification with deep convolutional neural networks". In: *arXiv preprint arXiv:1601.07255* (2016) (cit. on p. 16).

[Wu16b] Wu, Shangxuan; Chen, Ying-Cong; Li, Xiang; Wu, An-Cong; You, Jin-Jie and Zheng, Wei-Shi: "An enhanced deep feature representation for person re-identification". In: *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE. 2016, pp. 1–8 (cit. on pp. 16, 110).

[Wu17] Wu, Ancong; Zheng, Wei-Shi; Yu, Hong-Xing; Gong, Shaogang and Lai, Jianhuang: "RGB-infrared cross-modality person re-identification". In: *ICCV*. IEEE, 2017 (cit. on p. 124).

[Wu18] Wu, Yu; Lin, Yutian; Dong, Xuanyi; Yan, Yan; Ouyang, Wanli and Yang, Yi: "Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on p. 29).

[Xia16a] Xiao, Tong; Li, Hongsheng; Ouyang, Wanli and Wang, Xiaogang: "Learning deep feature representations with domain guided dropout for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1249–1258 (cit. on pp. 17, 108).

[Xia16b] Xiao, Tong; Li, Shuang; Wang, Bochao; Lin, Liang and Wang, Xiaogang: "End-to-end deep learning for person search". In: *arXiv preprint arXiv:1604.01850* 2 (2016) (cit. on pp. 29, 111, 113–117).

[Xia17] Xiao, Tong; Li, Shuang; Wang, Bochao; Lin, Liang and Wang, Xiaogang: "Joint detection and identification feature learning for person search". In: *Proc. CVPR*. 2017 (cit. on p. 94).

[Xia19] Xiao, Jimin; Xie, Yanchun; Tillo, Tammam; Huang, Kaizhu; Wei, Yunchao and Feng, Jiashi: "IAN: the individual aggregation network for person search". In: *Pattern Recognition* 87 (2019), pp. 332–340 (cit. on p. 97).

[Xie16] Xie, Junyuan; Girshick, Ross and Farhadi, Ali: "Unsupervised Deep Embedding for Clustering Analysis". In: *International Conference on Machine Learning (ICML)*. 2016 (cit. on pp. 103, 104).

[Xu17] Xu, Shuangjie; Cheng, Yu; Gu, Kang; Yang, Yang; Chang, Shiyu and Zhou, Pan: "Jointly attentive spatial-temporal pooling networks for video-based person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4733–4742 (cit. on p. 19).

[Yan16a] Yan, Yichao; Ni, Bingbing; Song, Zhichao; Ma, Chao; Yan, Yan and Yang, Xiaokang: "Person re-identification via recurrent feature aggregation". In: *European Conference on Computer Vision*. Springer. 2016, pp. 701–716 (cit. on p. 19).

[Yan16b] Yang, Luwei; Zhu, Ligen; Wei, Yichen; Liang, Shuang and Tan, Ping: "Attribute recognition from adaptive parts". In: *arXiv preprint arXiv:1607.01437* (2016) (cit. on p. 31).

[Yan16c] Yang, Yang; Liao, Shengcai; Lei, Zhen and Li, Stan Z: "Large Scale Similarity Learning Using Similar Pairs for Person Verification." In: *AAAI*. 2016, pp. 3655–3661 (cit. on p. 14).

[Ye15] Ye, Mang; Chen, Jun; Leng, Qingming; Liang, Chao; Wang, Zheng and Sun, Kaimin: "Coupled-view based ranking optimization for person re-identification". In: *International Conference on Multimedia Modeling*. Springer. 2015, pp. 105–117 (cit. on p. 22).

[Ye16] Ye, Mang; Liang, Chao; Yu, Yi; Wang, Zheng; Leng, Qingming; Xiao, Chunxia; Chen, Jun and Hu, Ruimin: "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing". In: *IEEE Transactions on Multimedia* 18.12 (2016), pp. 2553–2566 (cit. on p. 23).

[Ye17] Ye, Mang; Ma, Andy J.; Zheng, Liang; Li, Jiawei and Yuen, Pong C.: "Dynamic Label Graph Matching for Unsupervised Video Re-Identification". In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 (cit. on p. 94).

[Ye18] Ye, Mang; Wang, Zheng; Lan, Xiangyuan and Yuen, Pong C: "Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking." In: *IJCAI*. 2018, pp. 1092–1099 (cit. on p. 124).

[Yi14] Yi, Dong; Lei, Zhen; Liao, Shengcai and Li, Stan Z: "Deep metric learning for person re-identification". In: *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE. 2014, pp. 34–39 (cit. on p. 18).

[Yin17] Yin, Zhou; Zheng, Wei-Shi; Wu, Ancong; Yu, Hong-Xing; Wang, Hai and Lai, Jianhuang: "Learning a Semantically Discriminative Joint Space for Attribute Based Person Re-identification". In: *arXiv preprint arXiv:1712.01493* (2017) (cit. on p. 125).

[Yu16] 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: *arXiv preprint arXiv:1611.05603* (2016) (cit. on pp. 31, 64, 65).

[Yu18] Yu, Rui; Dou, Zhiyong; Bai, Song; Zhang, Zhaoxiang; Xu, Yongchao and Bai, Xiang: "Hard-aware point-to-set deep metric for person re-identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 188–204 (cit. on p. 19).

[Zaj05] Zajdel, Wojciech; Zivkovic, Zoran and Krose, BJA: "Keeping track of humans: Have I seen this person before?" In: *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE. 2005, pp. 2081–2086 (cit. on p. 2).

[Zei12] Zeiler, Matthew D: "ADADELTA: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701* (2012) (cit. on p. 44).

[Zha13] Zhao, Rui; Ouyang, Wanli and Wang, Xiaogang: "Unsupervised salience learning for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3586–3593 (cit. on p. 12).

[Zha14] Zhang, Ning; Paluri, Manohar; Ranzato, Marc'Aurelio; Darrell, Trevor and Bourdev, Lubomir: "Panda: Pose aligned networks for deep attribute modeling". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1637–1644 (cit. on p. 31).

[Zha17a] Zhang, Xuan; Luo, Hao; Fan, Xing; Xiang, Weilai; Sun, Yixiao; Xiao, Qiqi; Jiang, Wei; Zhang, Chi and Sun, Jian: "Alignedreid: Surpassing human-level performance in person re-identification". In: *arXiv preprint arXiv:1711.08184* (2017) (cit. on p. 124).

[Zha17b] Zhao, Haiyu; Tian, Maoqing; Sun, Shuyang; Shao, Jing; Yan, Junjie; Yi, Shuai; Wang, Xiaogang and Tang, Xiaoou: "Spindle net: Person re-identification with human body region guided feature decomposition and fusion". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1077–1085 (cit. on p. 22).

[Zha17c] Zhao, Liming; Li, Xi; Wang, Jingdong and Zhuang, Yueting: "Deeply-Learned Part-Aligned Representations for Person Re-Identification". In: *ICCV* (2017) (cit. on pp. 19, 22, 94).

[Zha18a] Zhang, Wei; Yu, Xiaodong and He, Xuanyu: "Learning bidirectional temporal cues for video-based person re-identification". In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.10 (2018), pp. 2768–2776 (cit. on p. 19).

[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 p. 33).

[Zhe09] Zheng, Wei-Shi; Gong, Shaogang and Xiang, Tao: "Associating Groups of People." In: *BMVC*. Vol. 2. 6. 2009 (cit. on p. 29).

[Zhe15a] Zheng, Liang; Shen, Liyue; Tian, Lu; Wang, Shengjin; Wang, Jingdong and Tian, Qi: "Scalable Person Re-identification: A Benchmark". In: *Computer Vision, IEEE International Conference on*. 2015 (cit. on p. 6).

[Zhe15b] Zheng, Liang; Shen, Liyue; Tian, Lu; Wang, Shengjin; Wang, Jingdong and Tian, Qi: "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1116–1124 (cit. on pp. 6, 29, 73, 114).

[Zhe16a] 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. 2016, pp. 868–884 (cit. on pp. 29, 93).

[Zhe16b] Zheng, Liang; Zhang, Hengheng; Sun, Shaoyan; Chandraker, Manmohan and Tian, Qi: "Person Re-identification in the Wild". In: *arXiv preprint arXiv:1604.02531* (2016) (cit. on pp. 29, 97, 111, 113, 115, 116).

[Zhe16c] Zheng, Wei-Shi; Gong, Shaogang and Xiang, Tao: "Towards open-world person re-identification by one-shot group-based verification". In: *IEEE transactions on pattern analysis and machine intelligence* 38.3 (2016), pp. 591–606 (cit. on p. 124).

[Zhe17a] Zheng, Liang; Huang, Yujia; Lu, Huchuan and Yang, Yi: "Pose invariant embedding for deep person re-identification". In: *arXiv preprint arXiv:1701.07732* (2017) (cit. on pp. 21, 87).

[Zhe17b] Zheng, Zhedong; Zheng, Liang and Yang, Yi: "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017 (cit. on pp. 6, 20, 22, 29, 73, 94).

[Zhe18] Zheng, Zhedong; Zheng, Liang and Yang, Yi: "A discriminatively learned cnn embedding for person reidentification". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1 (2018), p. 13 (cit. on pp. 19, 95).

[Zho17a] Zhong, Zhun; Zheng, Liang; Cao, Donglin and Li, Shaozi: "Re-ranking Person Re-identification with k-reciprocal Encoding". In: *CVPR*. 2017, pp. 1318–1327 (cit. on pp. 23, 94).

[Zho17b] Zhou, Zhen; Huang, Yan; Wang, Wei; Wang, Liang and Tan, Tieniu: "See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-Identification". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (cit. on p. 94).

[Zho18] Zhong, Zhun; Zheng, Liang; Zheng, Zhedong; Li, Shaozi and Yang, Yi: "Camera style adaptation for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5157–5166 (cit. on p. 20).

[Zhu13] Zhu, Jianqing; Liao, Shengcai; Lei, Zhen; Yi, Dong and Li, Stan: "Pedestrian attribute classification in surveillance: Database and evaluation". In: *Proceedings of the IEEE international conference on computer vision workshops*. 2013, pp. 331–338 (cit. on p. 37).

[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)*. IEEE. 2015, pp. 535–540 (cit. on pp. 20, 31).

[Zhu17] Zhu, Feng; Li, Hongsheng; Ouyang, Wanli; Yu, Nenghai and Wang, Xiaogang: "Learning spatial regularization with image-level supervisions for multi-label image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5513–5522 (cit. on p. 64).

[Zhu18] Zhu, Xiatian; Wu, Botong; Huang, Dongcheng and Zheng, Wei-Shi: "Fast open-world person re-identification". In: *IEEE Transactions on Image Processing* 27.5 (2018), pp. 2286–2300 (cit. on p. 124).

Own Publications

[1] Bäuml, Martin; Tapaswi, Makarand; Schumann, Arne and Stiefelhagen, Rainer: "Contextual Constraints for Person Retrieval in Camera Networks". In: *Advanced Video and Signal-Based Surveillance (AVSS), IEEE Ninth International Conference on*. IEEE. 2012.

[2] Schumann, Arne; Bäuml, Martin and Stiefelhagen, Rainer: "Person Tracking-by-detection with Efficient Selection of Part-detectors". In: *Advanced Video and Signal Based Surveillance (AVSS), 10th IEEE International Conference on*. IEEE. 2013.

[3] Schumann, Arne and Monari, Eduardo: "A Soft-biometrics Dataset for Person Tracking and Re-Identification". In: *Advanced Video and Signal Based Surveillance (AVSS), 11th IEEE International Conference on*. IEEE. 2014.

[4] Schumann, Arne: "Object Instance Recognition using Motion Cues and Instance Specific Appearance Models". In: *Video Surveillance and Transportation Imaging Applications*. Vol. 9026. International Society for Optics and Photonics. 2014, 90260S.

[5] Schumann, Arne and Stiefelhagen, Rainer: "Transferring Attributes for Person Re-Identification". In: *Advanced Video and Signal Based Surveillance (AVSS), 12th IEEE International Conference on*. IEEE. 2015.

[6] Schumann, Arne and Schuchert, Tobias: "Person Re-Identification in UAV Videos using Relevance Feedback". In: *Video Surveillance and Transportation Imaging Applications*. Vol. 9407. International Society for Optics and Photonics. 2015, 94070Z.

[7] Schumann, Arne and Schuchert, Tobias: "Deep Person Re-Identification in Aerial Images". In: *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*. Vol. 9995. International Society for Optics and Photonics. 2016, p. 99950M.

[8] Schumann, Arne; Gong, Shaogang and Schuchert, Tobias: "Deep Learning Prototype Domains for Person Re-Identification". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017.

[9] Sarfraz, M Saquib; Schumann, Arne; Wang, Yan and Stiefelhagen, Rainer: "Deep View-Sensitive Pedestrian Attribute Inference in an end-to-end Model". In: *British Machine Vision Conference (BMVC)*. 2017.

[10] Schumann, Arne and Stiefelhagen, Rainer: "Person Re-Identification by Deep Learning Attribute-Complementary Information". In: *Computer Vision and Pattern Recognition (CVPR) Workshop, IEEE Conference on*. IEEE. 2017.

[11] Schumann, Arne and Metzler, Jürgen: "Person Re-Identification Across Aerial and Ground-based Cameras by Deep Feature Fusion". In: *Automatic Target Recognition XXVII*. Vol. 10202. International Society for Optics and Photonics. 2017, 102020A.

[12] Schumann, Arne; Specker, Andreas and Beyerer, Jürgen: "Attribute-based Person Retrieval and Search in Video Sequences". In: *Advanced Video and Signal Based Surveillance (AVSS) Workshop, 15th IEEE International Conference on*. IEEE. 2018.

[13] Sarfraz, M Saquib; Schumann, Arne; Eberle, Andreas and Stiefelhagen, Rainer: "A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking". In: *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE. 2018.

[14] Schumann, Arne and Metzler, Jürgen: "Adapted Deep Feature Fusion for Person Re-Identification in Aerial Images". In: *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*. Vol. 10643. International Society for Optics and Photonics. 2018, p. 106430L.

[15] Specker, Andreas; Schumann, Arne and Beyerer, Jürgen: "An Interactive Framework for Cross-Modal Attribute-based Person Retrieval". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019.

Own Publications off Topic

[1] Saur, Günter; Krüger, Wolfgang and Schumann, Arne: "Extended Image Differencing for Change Detection in UAV Video Mosaics". In: *Video Surveillance and Transportation Imaging Applications*. Vol. 9026. International Society for Optics and Photonics. 2014, p. 90260L.

[2] Sommer, Lars; Nie, Kun; Schumann, Arne; Schuchert, Tobias and Beyerer, Jürgen: "Semantic Labeling for Improved Vehicle Detection in Aerial Imagery". In: *Advanced Video and Signal Based Surveillance (AVSS), 14th IEEE International Conference on*. IEEE. 2017.

[3] Coluccia, Angelo; Ghenescu, Marian; Piatrik, Tomas; De Cubber, Geert; Schumann, Arne; Sommer, Lars; Klatte, Johannes; Schuchert, Tobias; Beyerer, Juergen; Farhadi, Mohammad, et al.: "Drone-vs-Bird Detection Challenge at IEEE AVSS2017". In: *Advanced Video and Signal Based Surveillance (AVSS) Workshop, 14th IEEE International Conference on*. IEEE. 2017.

[4] Schumann, Arne; Sommer, Lars; Klatte, Johannes; Schuchert, Tobias and Beyerer, Jürgen: "Deep Cross-domain Flying Object Classification for Robust UAV Detection". In: *Advanced Video and Signal Based Surveillance (AVSSW) Workshop, 14th IEEE International Conference on*. IEEE. 2017.

[5] Sommer, Lars; Schumann, Arne; Müller, Thomas; Schuchert, Tobias and Beyerer, Jürgen: "Flying Object Detection for Automatic

UAV Recognition". In: *Advanced Video and Signal Based Surveillance (AVSS) Workshop, 14th IEEE International Conference on*. IEEE. 2017.

[6] Lyu, Siwei et al.: "UA-DETRAC 2018: Report of AVSS2018 & IWT4S Challenge on Advanced Traffic Monitoring". In: *Advanced Video and Signal Based Surveillance (AVSS) Workshop, 15th IEEE International Conference on*. IEEE. 2018.

[7] Acatay, Oliver; Sommer, Lars; Schumann, Arne and Jürgen, Beyerer: "Ensemble of Two-Stage Regression Based Detectors for Accuracte Vehicle Detection in Traffic Surveillance Data". In: *Advanced Video and Signal Based Surveillance (AVSS) Workshop, 15th IEEE International Conference on*. IEEE. 2018.

[8] Acatay, Oliver; Sommer, Lars; Schumann, Arne and Jürgen, Beyerer: "Comprehensive Evaluation of Deep Learning based Detection Methods for Vehicle Detection in Aerial Imagery". In: *Advanced Video and Signal Based Surveillance (AVSS), 15th IEEE International Conference on*. IEEE. 2018.

[9] Schumann, Arne; Sommer, Lars and Vogler, Max: "Onthology-based Masking Loss for Improved Generalization in Remote Sensing Semantic Retrieval". In: *Advanced Video and Signal Based Surveillance (AVSS), 15th IEEE International Conference on*. IEEE. 2018.

[10] Sommer, Lars; Schmidt, Nicole; Schumann, Arne and Beyerer, Jürgen: "Search Area Reduction Fast-RCNN for Fast Vehicle Detection in Large Aerial Imagery". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3054–3058.

[11] Zhu, Pengfei et al.: "VisDrone-DET2018: The Vision Meets Drone Object Detection in Image Challenge Results". In: *IEEE European Conference on Computer Vision (ECCV) Workshop*. IEEE. 2018.

[12] Nie, Kun; Sommer, Lars; Schumann, Arne and Beyerer, Jürgen: "Semantic Labeling Based Vehicle Detection in Aerial Imagery". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018.

[13] Sommer, Lars; Schumann, Arne; Schuchert, Tobias and Beyerer, Jürgen: "Multi Feature Deconvolutional Faster R-CNN for Precise Vehicle Detection in Aerial Imagery". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018.

[14] Sommer, Lars; Steinmann, Lucas; Schumann, Arne and Beyerer, Jürgen: "Systematic Evaluation of Deep Learning Based Detection Frameworks for Aerial Imagery". In: *Automatic Target Recognition XXVIII*. Vol. 10648. International Society for Optics and Photonics. 2018, p. 1064803.

[15] Valev, Krassimir; Schumann, Arne; Sommer, Lars and Beyerer, Jürgen: "A Systematic Evaluation of Recent Deep Learning Architectures for Fine-grained Vehicle Classification". In: *Pattern Recognition and Tracking XXIX*. Vol. 10649. International Society for Optics and Photonics. 2018, p. 1064902.

[16] Schumann, Arne; Sommer, Lars; Müller, Thomas and Voth, Sascha: "An Image Processing Pipeline for Long Range UAV Detection". In: *Emerging Imaging and Sensing Technologies for Security and Defence III; and Unmanned Sensors, Systems, and Countermeasures*. Vol. 10799. International Society for Optics and Photonics. 2018, 107990T.

[18] Steinmann, Lucas; Sommer, Lars; Schumann, Arne and Beyerer, Jürgen: "A Fast and Lightweight Person Detector for Unmanned Aerial Vehicles". In: *European Signal Processing Conference (EUSIPCO) Workshop*. 2019.

[19] Azimi, Seyed Majid; Henry, Corentin; Sommer, Lars; Schumann, Arne and Vic, Eleonora: "Skyscapes - Fine-grained Unterstanding of Aerial Scenes". In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2019.

[17] Ringwald, Tobias; Sommer, Lars; Schumann, Arne and Beyerer, Jürgen: "UAV-Net: A Fast Aerial Vehicle Detector for Mobile Platforms". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*. IEEE. 2019.

List of Figures

List of Tables

Acronyms

ACR	Attribute-Complementary Re-Id
AP	Average Precision
CCTV	Closed Circuit Television
CMC	Cumulative Matching Characteristic
CNN	Convolutional Neural Network
DLDP	Deep Learning Domain Prototypes model
DPM	Deformable Parts Model object detector
ECN	Expanded Cross-Neighborhood re-ranking
FoV	Field of view
GAN	Generative Adversarial Network
LMNN	Large-Margin Nearest-Neighbor Learning
LOMO	Local Maximal Occurrence feature descriptor

LSTM	Long Short-Term Memory network
mAP	Mean Average Precision
MLP	Multi Layer Perceptron
MRF	Markov Random Field
NN	Nearest Neighbor
PCA	Principal Component Analysis
PGA	Pose-guided Attribute Attention model
PRW	Person Re-Identification in the Wild dataset
PSAE	Pose-Sensitive Attribute Embedding
PSE	Pose-Sensitive Embedding
re-id	Person Re-Identification
ReLU	Rectified Linear Unit
RNN	Recurrent Convolutional Neural Network
ROI	Region of Interest
SDALF	Symmetry Driven Accumulation of Local Features
VGG	Oxford University Visual Geometry Group