Semantic Alignment in Multiple Stages of Networks for Person Re-ID

Towards Generalizable Models

Ravi Autar

TU Delft

Semantic Alignment in Multiple Stages of Networks for Person Re-ID

Towards Generalizable Models

by

Ravi Autar

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on Tuesday November 12, 2019 at 09:00 AM.

Student number:	4361172	
Project duration:	February 4, 2019 – November 12, 2019	
Thesis committee:	Dr. Hayley Hung,	TU Delft
	Dr. Jan van Gemert,	TU Delft
	Dr. Laura Cabrera Quiros,	TU Delft
	Dr. Henri Bouma,	TNO
	MSc. Arthur van Rooijen,	TNO

Abstract

Person re-identification (re-ID) is a task that aims to associate the same people across different cameras. One of the many important problems a person re-ID system has to address in order to achieve good performance is the feature misalignment problem. Past research has attempted to address this problem by using attention networks, pose-estimation modules, or semantic segmentation networks. However, they all eventually tend to pool these features to a single feature embedding, thereby not distinguishing regions with different semantic meanings such as the head, torso, and lower body. Most approaches also do not make use of all the information available throughout multiple layers (stages) of the feature extractor. Furthermore, although these additional features are used to provide extra information to the re-ID network during training, they do not take into account the importance of different regions of the image due to, for example, occlusion. To circumvent these problems, we propose a network that is capable of extracting regional feature embeddings that are associated with specific body parts of an identity, i.e., head, upper-body, lower-body, shoes, and foreground image. We extract these features from multiple stages of a feature extractor using a semantic-segmentation module. We then use multi-branch learning to ensure that these features are independently optimized by introducing separate modules (branches) for each regional feature embedding. To increase the robustness of the model, we also propose a novel testing strategy that makes use of the importance and visibility of specific body parts in both the query and gallery images in order to calculate a ranking list. Finally, to address the current lack of datasets that contain images from overhead face-down cameras, we introduce a new dataset named MatchNMingle-reID. Because of the viewpoint of the cameras, this dataset presents unique challenges that are not seen in current datasets and opens possibilities to create more generalizable models that can effectively address the feature misalignment problem.

Contents

1 Introduction

1.1. Background

Person re-identification (re-ID) is a critical task in intelligent video surveillance, aiming to associate the same people across different cameras. Person re-ID is crucial for tasks such as tracking down suspects or missing people (elderly, children, etc.) and improving general public safety. For these reasons, governments and companies have invested many resources in increasing the number of surveillance cameras in areas such as airports, city centers, shopping malls, and many other public and non-public places. This type of digital surveillance generates massive amounts of data, making it impossible for human operators to process. This makes it necessary to automate the task of person re-ID by leveraging the power of computer vision.

Most surveillance cameras record images at low resolutions and typically capture a large field of view. This consequently leads to lower quality images being available of persons who may be of interest. Therefore, unlike facial recognition, the task of person re-ID involves retrieving images of the same identity using information about the face, but also of the clothing, hair, height, accessories, etc. Person re-ID systems require two essential components to achieve this goal, i.e. (1) a method for extracting features from images and (2) a similarity measure between the features of two different images. The quality of a person re-ID system strongly depends on how well images are expressed in terms of features. Therefore, many researchers mainly focus on improving the feature-extraction aspect of the person re-ID system.

In the past, researchers built re-ID systems using so-called “hand-crafted” features, which are clever ways to represent the image of a person. However, these methods have recently been surpassed by feature extraction through Deep Neural Networks (DNNs), which consequently are also the main topic of interest in our work. Person re-ID research primarily focuses on improving feature extractors to produce expressive and more discriminative representations of input images. These representations are then used to measure the similarity to other images. As a result, person re-ID is formulated as a ranking problem: given the query image of a person, the model needs to rank all gallery images based on their similarity to the query image.

1.2. Problem and Motivation

To understand the challenges that person re-ID models face in practice, researchers train and test their models on various benchmark datasets. These datasets are created to replicate the challenges faced in practice, which are caused by changes in visual appearance such as change of viewpoint, different image resolutions, pose variation, background clutter, and occlusion across and within datasets. Change of viewpoint is apparent across different datasets, where the camera is placed on a slightly different viewing angle. Pose variation is common since people move between different frames and cameras. Also, images within and across datasets tend to have various resolutions, while the final layer of most DNNs often consists of a fully connected network which requires a fixed size input. Finally, models do not tend to generalize well if the degree of difficulty increases across datasets in terms of background clutter and occlusions. These challenges are common across different datasets and give rise to the problem known as *feature misalignment* [27, 28, 37, 38, 48, 58]. Feature misalignment occurs when a particular region in the feature maps of two different images does not encode the same semantic information e.g., features of the head of a person are compared with the information from the upper-body. This consequently leads to an inaccurate estimation of similarity scores between images. Therefore, in order to be effective across different datasets and in practice, a model has to be invariant to most of these challenges.

One example where changes in viewpoint can lead to problems with the alignment of features is when we consider surveillance images taken from down-facing overhead cameras. Images taken from overhead cameras can pose unique challenges that are not present in current datasets because they always contain frontal view images of people. Images from an overhead viewpoint, cause a person in the middle of the frame to only be identifiable by the top of their heads and shoulders. In this viewpoint, the majority of people in any frame mainly have their heads and upper-bodies exposed to the camera. Finally, extracted bounding boxes of people appear to be rotated depending on the location of the person in the frame. All these points emphasize the feature misalignment problem of DNNs mentioned earlier. Despite the widespread use of this type of surveillance in practice, we have found that researchers have currently made no effort to investigate these scenarios, nor is there any dataset available to facilitate this work. Bottom-right of Figure 1.1 shows examples of images seen with the overhead viewpoint, along with examples of occlusion, pose variation and change in resolution.

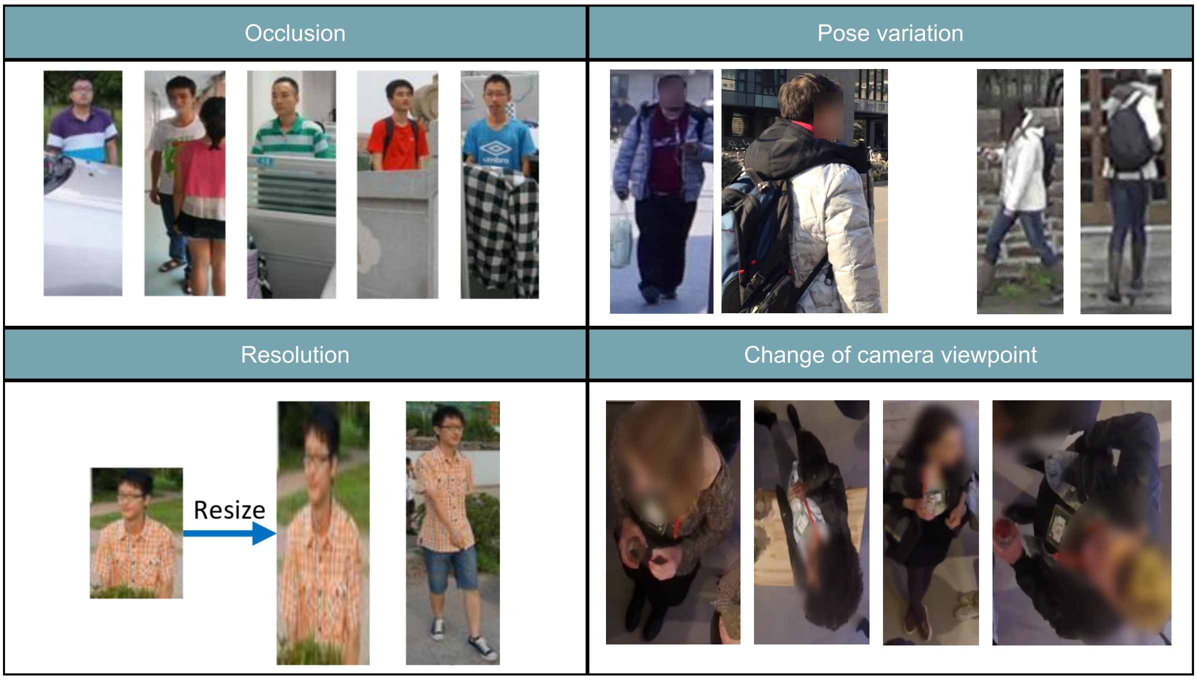

Figure 1.1: Examples of challenges related to the feature alignment problem. The left column shows examples of occlusion and change of resolution (Image source: *He et al.* [19]). Top right shows an example of pose variation within a dataset (Image source: *Wei et al.*[49]). Bottom right shows examples of images taken from a very different viewing angle than traditional datasets (Image source: *Cabrera et al.* [2]).

Researchers have tackled challenges related to feature misalignment with varying degrees of success. To alleviate the problem of different input image resolutions, most researchers have resorted to reshaping the input images. However, *He et al.* [19] empirically show that this results in unwanted deformations, which further degrades the performance of the model. To solve problems with occlusion and background clutter, researchers look at various data augmentation techniques such as random occlusion and random backgrounds. To achieve view invariance, many researchers have incorporated feature maps extracted from multiple semantic levels of their network to construct a final feature embedding using various fusion techniques [4, 47]. These semantic levels are also referred to as stages. To solve the feature misalignment problem in general, researchers have tried to incorporate attention networks, pose-estimation modules, and semantic masks (i.e., masks for isolating certain body parts) into their models, often with improved results [22, 26]. Similarly, others have introduced multiple branches into their network. Each branch is an independent learning network that is designed to learn a feature representation for a particular body part.

While the current approaches perform well on benchmarks datasets, many do not take into account some fundamental problems. Most approaches tend to pool information in a way that does not preserve the semantic discrepancy between different image regions, nor do they take into account the importance of different regions during testing. Furthermore, many multi-stage approaches pool intermediate feature maps at a certain point, while ignoring the feature misalignment problem. On the other hand, many multi-branch approaches do not use the information from earlier stages of their network, therefore missing out on more fine-grained information encoded in these layers. Finally, many approaches assume that people in images are always standing upright, while down-facing surveillance cameras capture images of people in various orientations. This further increases the problem of feature alignment.

In our work, we argue that we can improve person re-ID models by enabling them to effectively learn features for regions of images which have different semantic meaning. We also believe that by taking into account the semantic discrepancy between different image regions and weighting in the importance of these regions during testing, we can further improve upon current approaches. We argue that these two approaches are required to address the feature alignment problem effectively.

1.3. Contributions

We propose an approach to tackle the feature alignment problem in particular. While others have proven the usefulness of multi-stage training, we argue that a model can further benefit from features in these intermediate layers by taking into account the semantic discrepancy in different regions of the image. Therefore, we hypothesize that by incorporating a trained semantic-segmentation network to multiple stages of the feature extractor, we can extract more discriminative features that are useful for person re-ID. We also argue that we can further improve the discriminative power of our network by introducing multiple branches to our network, each of which is trained to extract features associated with particular body parts of a person visible in an image. Finally, we argue that while learning features for different regions is necessary, their relevance in terms of visibility should also be taken into account during test time.

To validate our hypotheses, we propose a model with two novel approaches, one during training and the other during testing. Our first novel approach incorporates a semantic segmentation network to multiple stages of the feature extractor. The segmentation network is trained to extract semantic masks for individual human body parts, i.e., the head, upper-body, lower-body, shoes, and the foreground image. In order to make efficient use of the features extracted from these earlier layers, we make use of fully connected layers to combine the features across multiple stages while maintaining the segmentation of the different body parts. The result of this is a multi-branch network, where each branch is responsible for learning the feature embedding of a particular body part across different stages of the base network. This multi-branch approach ensures us that the extracted features for each body part get optimized through independent supervision, without any potential negative mutual influence due to their semantic discrepancy.

Our second contribution focuses on making efficient use of features learned by our proposed network. Therefore, during testing, instead of computing a single similarity score between two images as most current approaches do, we compute a similarity score associated with each body part. We then combine them into a single similarity score by weighting the importance of the body parts. The importance of each part is determined by their visibility in both the query and gallery image.

In addition to a novel architecture, we also contribute by addressing the current lack of datasets that contain images from face-down overhead cameras. We do this to assess the performance of current methods on a dataset with significant viewpoint changes. Therefore, we introduce a new person re-ID dataset called MatchNMingle-reID. Created by *Cabrera et al.* [2], the original MatchNMingle dataset contains video recordings of people using face-down overhead cameras. To the best of our knowledge, the person re-ID community has not yet done any research on the cases where only overhead surveillance images are available. Therefore, with our work, we hope to bridge this gap by also presenting preliminary results on this new dataset.

In summary, our work can be grouped into three main contributions. (1) We propose a novel architecture that is capable of learning features associated with different human body parts throughout multiple stages of a feature extractor using a multi-branch approach. (2) We propose a method to make use of these features during testing by appropriately weighting the importance of particular body parts. (3) We introduce a new dataset named MatchNMingle-reID to the person re-ID domain, with unique challenges in terms of viewpoint.

The outline of this thesis is as follows: Chapter 2 touches upon related work and their contributions in attempting to solve the multitude of problems encountered in person re-ID. Chapter 3 discusses our model architecture and design choices in greater detail. Chapter 4 documents the experiments and results we obtain with different configurations of our model. We discuss the creation and preliminary tests on the MatchNMingle-reID dataset in Chapter 5, and we form our conclusion in Chapter 6.

2 Related Work

Person re-ID is formulated as a ranking problem: given a probe image of a person, the model needs to rank all gallery images based on their similarities to the probe image. The ranking performance is dependent on the quality of feature embedding of an image. There are various challenges present in different person re-ID datasets and real-life, which impair the performance of person re-ID models. In order to learn expressive and discriminative feature embeddings that are robust to these challenges, past research has attempted a variety of different approaches. Most of the research falls under the category of improving feature extractors. Another approach is to enrich the training dataset with new examples such that the produced feature embedding is robust to a variety of different scenarios. We discuss various different techniques for both feature extraction (Section 2.1) and data augmentation (Section 2.2). Additionally, we give an overview of the performances of the most successful techniques on three widely used benchmark datasets (Market-1501 and DukeMTMC-reID) in Section 2.3.

2.1. Feature Extraction

The ranking performance in person re-ID heavily relies on the quality of the feature embedding extracted from an image, which is usually learned from data. Before deep learning methods dominated the re-ID research community, so-called “hand-crafted” algorithms were used to learn part or local features. Images of people can be partitioned into horizontal stripes to extract color and texture features [15, 32, 33, 57]. More sophisticated strategies include dividing the images into several triangles for part feature extraction [12]. *Cheng et al.* [7] employ pictorial structure to parse people into semantic parts. *Das et al.* [10] use HSV (hue, saturation and value) histograms on the head, torso and legs to capture spatial information.

Nowadays, the majority of person re-ID research investigates techniques to extract more discriminative feature embeddings using Convolutional Neural Networks (CNNs) [24] as feature extractors. Common feature extractors for images include (but are not limited to) XceptionNet [8], ResNet50 [18] and DenseNet [20], and have also shown their effectiveness in the domain of person re-ID [21, 22, 26, 44, 45, 47]. Therefore, researchers often take these (pre-trained) feature extractors as a starting point for their work and refer to them as their base feature extractor. In the literature, we can group the approaches that researchers take to improve the base feature extractors into three categories: multi-stage, multi-branch, and selective feature extraction. This section discusses research performed in each of these categories separately. It is important to note that the categories mentioned above are not strictly exclusive; some research may fall into more than one category and are, therefore, also discussed in multiple sections.

2.1.1. Multi-Stage Feature Extraction

The common approach for person re-ID is to run images through a DNN feature extractor and use the coarse-resolution and semantic embeddings from the last layer to look the image up in a database/gallery. Although the high-level features extracted in the last layer of DNNs are useful in forming abstract concepts for object recognition, they discard low-level signals like color and texture, which are important clues for person re-ID. Features extracted at such coarse resolutions are not capable of encoding fine-level details such as patterns on clothes, facial features, subtle pose differences, etc. [45, 47]. *Chang et al.* [4] also argued that multi-stage feature extraction can be beneficial in extracting view-invariant features. These arguments suggest that person re-ID will benefit from fusing information across multiple layers of a feature extractor. This is generally done by splitting a DNN feature extractor into multiple stages and using the intermediate feature representations in these stages to train the network and to construct a final feature embedding for similarity estimation [29, 37, 45, 47].

Fusion There are several ways to effectively make use of embeddings across multiple stages of a network. One of the approaches is to create a fusion embedding of the multi-stage features. DARENet [47] splits the ResNet50 [18] model into four stages and processes the intermediate feature maps to the same dimensions using global average pooling (GAP) and fully connected (FC) layers. These intermediate feature embeddings are then fused to a final feature embedding using the weighted sum of learnable parameters. Another approach is to compute multiple similarity scores using each embedding and to compute their weighted average [6]. The MLFN [4] network fuses features from multiple stages using a specialized module called the Factor Selection Module (FSM). The FSM dynamically selects features from multiple stages based on their activations and feeds a compact representation of the features in every stage directly to the final layer. Here they are aggregated together with the features from the final stage to create a fusion embedding. Other approaches concatenate the features across multiple stages to construct a new feature embedding used for classification [17, 37]. We also make use of concatenated features in our model. However, instead of creating a single concatenated feature embedding that represents the entire image, we first segregate features associated with different body parts at every stage. This approach gives us multiple feature embeddings, each of which uniquely represents a body part.

Specialized modules Several researchers have included specialized modules to make use of the features maps in multiple stages effectively. Modules have been introduced to align feature maps in multiple stages [19, 28, 37, 38, 48], account for background clutter [39] and improve the measurement of similarity between input images [17]. The ReSAnet [45] network introduced the Spatial Attention (SA) module in each stage of their network, before performing a global average pooling (GAP) operation. The SA module is a parameter-free module that assigns importance to different spatial positions of feature maps. The importance of a spatial position is computed based on the total intensity of the activations along the corresponding channel. Therefore, it regularizes the GAP in a complementary way where all learned filters compete with each other to become more important. After performing GAP, the features undergo a regular classification procedure using an FC layer. In addition to SA modules in multiple stages of the network, ReSAnet also uses the Parts Convolutional Baseline (PCB) module [41] on the feature maps of the final stage. The feature embedding at the last stage of the backbone network is then used for distance calculation. Due to their recent successes in person re-ID, we also use the SA and PCB modules in our work.

Deep supervision Researches discussed so far compute a fusion feature embedding, which has a single loss term associated with the final feature embedding. However, some research also assigns loss terms to the feature embeddings in intermediate stages, which are used to train the network by making predictions. This type of multi-stage training is also referred to as deep supervision [25, 51], and has also been shown to be effective for person re-ID [45, 47]. This technique can lead to more discriminative intermediate features. Due to their success in recent research [45, 47], we also make use of this in our network. However, instead of training features of a single stage, we train features associated with specific body parts across multiple stages of the network.

2.1.2. Multi-branch Feature Extraction

The feature misalignment problem occurs when the same region in different images does not encode the same semantic information. To tackle this problem, several researchers have taken a multi-branch learning approach. In this approach, the re-ID network gets subdivided into multiple branches, where each branch is responsible for learning a particular feature. The idea is that features can be learned independently without possible negative mutual influences due to semantic discrepancy during training.

One approach for multi-branch learning trains a separate module to complement the main feature extractor network by extracting specific attributes present in images, such as gender, age, and accessories [46]. However, this approach assumes that annotated attribute labels are available in the dataset. Another approach is to train a network to extract features associated with different body parts. Several approaches do this by using attention maps and incorporating them into their main feature extractor [28, 52]. Although attention maps can give an indication of important regions in feature maps, it does not provide any semantic information of these regions. One approach introduces a regularization term between the different branches to ensure that the extracted features are diverse from one another, regardless of the actual semantic meaning of these body parts [26]. Finally, one approach makes use of regional masks extracted using a semantic-segmentation module to independently construct features associated with the head, torso, and lower-body of a person [43]. In our work, we also use a semantic-segmentation network to guide our feature extractor to extract features from different body parts. However, unlike previous research which extracts features from a single stage, we extract these regional features across multiple stages of our base network.

Extracted features from different branches are used in various ways. One approach which independently extracts attribute features smoothly fuses this information with the global feature branch using an encoder-decoder network [46]. Another approach trained the global feature extractor and specialized branches separately [28, 43]. A loss term is associated with the global feature extractor, and another loss term is used to train the specialized branches. The features from the global feature extractor and the specialized branches are then concatenated for similarity estimation. While we also train our branches separately, we do not simply concatenate the feature embeddings for similarity estimation. Instead, we use an approach to take into account whether or not certain body parts are visible in the image. We then compute separate similarity estimations for each body part, before weighting them based on their visibility to compute the final similarity score.

The multi-branch approach discussed in this section is powerful because it has the potential to solve the feature misalignment problem and provides a good method to improve the performance of a model due to its capability to independently optimize each branch for extracting certain features. Despite the successes of multi-branch approaches, many researchers tend to train independent networks for extracting a particular feature, but show no explicit use of features extracted from earlier stages of the model. We argue that a multi-branch approach can benefit from incorporating more fine-grained features encoded in earlier stages of the network, thereby improving their performance.

2.1.3. Selective Feature Comparison

Another approach that aims to solve the feature alignment problem is to use selective feature extraction. Factors like background clutter, pose, and camera point of view variations hinder the process of extracting robust and discriminative representations, hence preventing different identities from being successfully distinguished. Several approaches for improving representation learning are based on bounding box part detection to extract local features from human body parts. However, they suffer from the low resolution of further dividing the input image to capture valuable features. The idea is to help a re-ID model focus more on the important regions of an image with the use of attention networks, pose estimation networks, or semantic/mask segmentation networks.

Several approaches demonstrate the effectiveness of using attention maps in their base model for feature extraction [31, 63]. In one approach, the network learns multiple spatial attention models [26]. It employs a diversity regularization term to ensure multiple models do not discover the same body part, by penalizing overlapping body parts detected by these modules. Another approach is to infuse the information from pose estimation models to the global feature extractor [36, 40, 54, 56]. However, these methods are all based on rigid body regions, which cannot accurately localize human body regions. One approach solves this by learning attention maps that are guided by pose estimation [52]. Here, non-rigid body regions are obtained based on the connectivity between human joints.

Recent work has also demonstrated improved performances by directly parsing different human body parts using semantic segmentation and integrating this information into the global feature extractor [22, 43]. *Kalayeh et al.* [22] argue that human semantic parsing is a better alternative to bounding boxes due to its pixel-level accuracy and capability of modeling arbitrary contours. The proposed work trains an Inception-V3 [42] semantic segmentation network to segment an image into five classes (fore-ground, head, upper, lower-body, and shoes). A different base feature extractor is used to extract features from the original image. These output activations are then pooled multiple times using the information from the semantic module (one pooling per semantic class) to extract information from the regions which contained one of the five classes. Due to the recent success of this approach, we chose to use a similar approach to extract features from different body parts. In contrast to the approaches mentioned above, we integrate our semantic segmentation network in multiple stages of the network, instead of only the final stage.

2.2. Data Augmentation

Increasing the size and diversity of training data is an effective way to make models more generalizable, which is why researchers create new and more massive datasets [49]. However, a technique called data augmentation allows models to generalize and perform better on existing datasets. Data augmentation techniques are useful in diversifying the training data such that a person re-ID model can train on new scenarios which are not encountered in the original dataset. Different approaches have shown the effectiveness of data augmentation in improving the robustness of a model against occlusion, background clutter, and illumination/color variations. Conventional approaches [23] include cropping, mirroring, warping, and jittering a training image. These approaches are also applicable in person re-ID for the same purpose. In our work, we have experimented with the cropping and mirroring data augmentation techniques as they have shown to be effective in past research [34, 41, 45, 47].

Apart from generic data augmentation techniques, research has also focused on data augmentation, which is specially designed for the context of person re-ID. Random Erasing [47, 61] is a technique used to make models robust to occlusion by randomly generating occluded training samples. This approach is extended to generate samples that are adversarial to the re-ID model by convolving a black square over the original image and finding which occlusion degrades the performance of the model [21]. Another approach increases the robustness of person re-ID to background similarities and variations [43]. This is done by randomly replacing the background of images in the training data with backgrounds taken from other surveillance images. Finally, to increase the variety and size of the training data, past research [11, 30, 49] made use of Generative Adversarial Networks (GANs) [14], specifically the CycleGAN [9].

Inspired by data augmentation techniques discussed in this section, we implement a new data augmentation technique for the MatchNMingle-reID dataset. Images of people in this dataset can appear in arbitrary orientations due to the viewpoint. Therefore, for this dataset, we introduced random rotations, which rotates the image of a person around its center at an arbitrary angle. This data augmentation technique is not possible in traditional datasets since people in these datasets always appear in an upright standing position.

2.3. Performance Comparison

In this section, we will compare the most successful approaches based on their performances on two widely known benchmark datasets for person re-ID: Market-1501 and DukeMTMC-reID. Throughout our work, we make use of two widely used standard evaluation metrics for person re-ID: Rank-1 Cumulative Matching Characteristic accuracy (R1) [16] and mean Average Precision (mAP) [55]. Therefore, we also make use of these metrics for the performance comparison in this section. Because person re-ID is formulated as a ranking problem, these evaluation metrics indicate the quality of the ranking. The R1 score indicates the probability that a query identity appears at the top of the rank list. This metric does not consider the total number of ground truth images in the gallery. For mAP, the idea is that a perfect re-ID system should be able to return all true matches of a particular class and therefore does consider all the ground truth images in the gallery.

We present the results of the best performing techniques in Table 2.1. By looking at the performance of these approaches, we see that the ReSAnet network proposed by *Wang et al.* [45] outperforms other techniques on the Market-1501 and DukeMTMC-reID datasets in both the mAP and R1 score. In this method, the network is trained using a parts-based classifier [41] and multi-level features, which are regularized using their novel Spatial Attention module, thereby achieving a mAP score of 91.7% and 85.9% on Market-1501 and DukeMTMC-reID datasets respectively.

	Market-1501			DUKEMTMC-reID		
Technique	**mAP (%)**	**R1 (%)**	**R5 (%)**	**mAP (%)**	**R1 (%)**	**R5 (%)**
Chen et al.[6]	81.6	93.5	97.7	69.5	84.9	92.3
He et al.[19]	82.5	92.7	**96.9**	66.4	80.7	88.5
Xu et al.[52]	82.96	88.69	-	59.25	76.84	-
Huang et al.[21]	83.3	88.66	-	78.19	84.11	-
Wang et al.[47]	86.7	90.9	-	80	84.4	-
Kaleyah et al.[22]	90.96	94.63	96.82	84.99	**88.96**	93.27
Wang et al.[45]	**91.7**	**94.7**	-	**85.9**	89.0	-

Table 2.1: Comparison of current state-of-the-art approaches for the Market-1501 and DukeMTMC-reID datasets.

The approach introduced by *Kalayeh et al.* [22] makes use of semantically segmented masks of 5 body parts and fuses them to construct a discriminative feature embedding at the final stage of their network. We can also observe that the results of this approach are comparable to that of *Wang et al.* [45], scoring a mAP of 90.96% and 88.96% on the Market1501 and DukeMTMC-reID datasets, respectively.

Wang et al. [47] makes clever use of the salient features available in lower semantic levels by fusing them to form a final feature embedding and achieves a mAP score of 86.7%, 80.0% on the Market1501 and DukeMTMC-reID datasets respectively. Other methods that achieve competitive results and seem to improve compared to previous state-of-the-art approaches use data augmentation techniques to account for background clutter and occlusion [21, 62], the best of which achieve a mAP of 83.3% and 78.19% on Market-1501 and DukeMTMC-reID respectively. Finally, the method proposed by *Li et al.* [26] makes use of a multi-branch approach to learning unique discriminative features automatically.

3 Method

Most existing methods have demonstrated their effectiveness on benchmark datasets, but still fail to generalize well across different datasets, which is evident from their degraded performances. The problem we aim to tackle in our work is the feature misalignment problem since this captures multiple other fundamental problems, some of which have not received much attention in past research. We do this by introducing a novel architecture with two concepts that are central to our work: multi-stage and multi-branch learning. Multi-stage refers to intermediate representations of feature maps across the depth of a DNN. Here it is useful to think of it as splitting a feature extractor into several parts (or stages) and making use of the intermediate feature maps. Multi-branch learning refers to the independent optimization of features by utilizing a smaller network.

Our model is capable of independently learning region-specific feature embeddings across multiple stages of a DNN, hereafter referred to as a base network. Using a semantic segmentation network (presented in Section 3.1), we extract five regional masks of the person, i.e., head, upper-body, lower-body, shoes, and foreground, hereafter referred to as semantic masks. These regional masks are then used in multiple stages of our base network (detailed in Section 3.2), to extract region-specific features from the feature maps in a particular stage. We introduce fully connected layers to learn a feature embedding for each body part visible in an image (Section 3.3). The latter is what we refer to as our multi-branch approach.

3.1. Semantic Segmentation Network

The first aspect of our model consists of a semantic-segmentation module, which is capable of extracting different regions of the human body. Improving on the approach taken by *Kalayeh et al.* [22], we employ the DeepLabV3+ model instead of Xception V3 and train it on the **Look into Person (LIP)** dataset [13] to extract five different regions from a given image: head, upper-body, lower-body, shoes, and foreground. The latter is the aggregate of the previous four. The reason for using semantic segmentation as opposed to bounding boxes over the desired regions is because of the superiority of the former in terms of pixel-level accuracy.

As opposed to the approach taken by *Kalayeh et al.* [22], our semantically segmented image is dependent on the height and width dimension of the input image, whereas the former extracts a fixed 30x30 semantic mask for each body part. The reason for doing this is that we want to downscale the segmentation mask using bi-linear interpolation to match whatever dimensions we obtain in different stages of the base network. This prevents us from having to re-scale our image or feature maps more along a particular axis than the other, which, as argued by *He et al.* [19], significantly degrades the performance due to undesired deformations of the source image.

Suppose an arbitrary image is of dimension ($w \times h \times 3$). With the semantic-segmentation network, we extract segmentation masks which have dimensions of ($w \times h \times 5$), where the depth channel denotes the five different semantic regions that we extract from the images. These masks are then downscaled along the height and width dimension using bi-linear interpolation, such that they match the dimensions of the feature maps at different stages of the base network. A representation of such a network can be seen in Figure 3.1 and some examples of semantic masks can be seen in Figure 3.2.

These segmentation masks are then used in our main re-ID network to filter feature maps based on the regions they represent. Therefore, the resolution of the down-scaled semantic masks is dependent on the spatial resolution of the feature maps at a particular level of the base network.

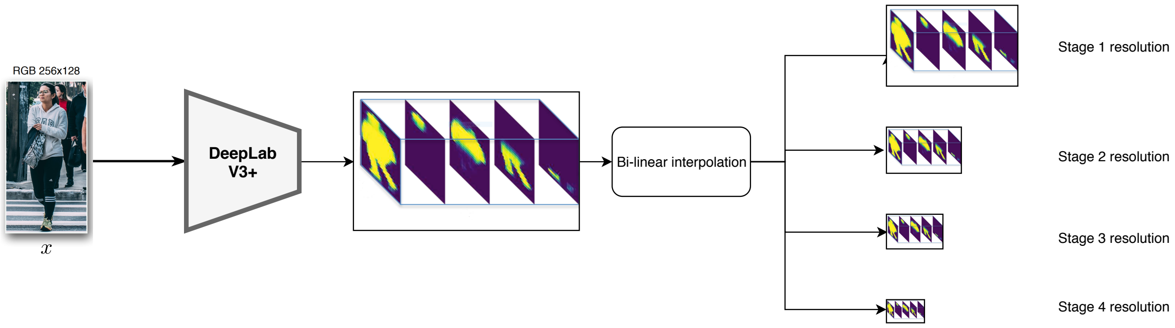

Figure 3.1: Semantic-segmentation module used to create masks for different stages of the network. The mask of the initial image is first computed and resized to resolution of different stages using bi-linear interpolation.

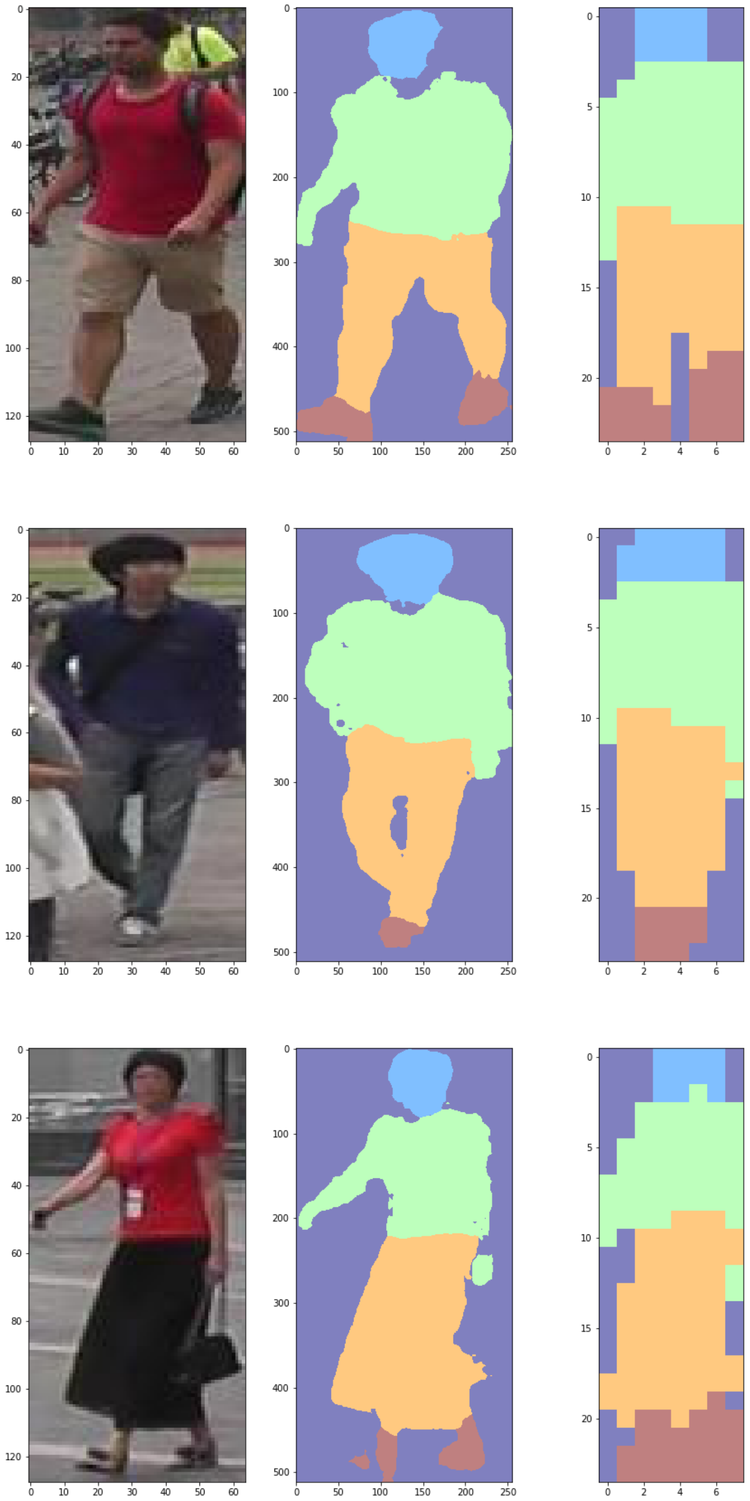

Figure 3.2: Examples of masks using semantic segmentation. From left to right these correspond to the original images, the mask of the original image and masks with the resolution of the feature maps at the last stage of the base network (24x8).

3.2. Multi-Stage Model

3.2.1. Semantic Segmentation Variant

Several different studies [4, 45, 47] have shown the effectiveness of using features from multiple levels of the base network on increasing the re-ID performance. On the other hand, *Kalayeh et al.* [22] have documented on the advantages of using semantic segmentation masks on the feature maps of the last stage of the re-ID network and train it to extract more discriminative features. However, the study conducted by *Wang et al.* [47], noted that the features of the final layer of their base network achieved a higher error rate than the features from the previous layer. A possible explanation for this is that the features in the final layer are too high level and lose their discriminative capability. Using this insight and because the spatial resolution of the feature maps in the final layer is relatively small, we argue that by performing semantic segmentation on these feature maps will remove too much information, thereby further degrading their discriminative capabilities.

To alleviate the aforementioned problems, we constructed a network that uses our trained semantic-segmentation module in multiple stages of the base network. This network is referred to as **Architecture 1A** for which a representation can be seen in Figure 3.3. Our model consists of a base network for which we have used ResNet50 due to its success in previous person re-ID research. Inspired by the work of *Wang et al.* [45], we split the ReSAnet network up in 4 stages, each stage extracting intermediate feature maps of the input image. This multi-stage approach is highlighted in light orange color in Figure 3.3. These intermediate feature representations are then passed through the Spatial Attention module [45] as a regularization step. As opposed to *Wang et al.* [45], who pass the maps through a GAP and FC layer at this point, we integrate the semantic information from the semantic-segmentation module using the Semantic Filtering (SF) module.

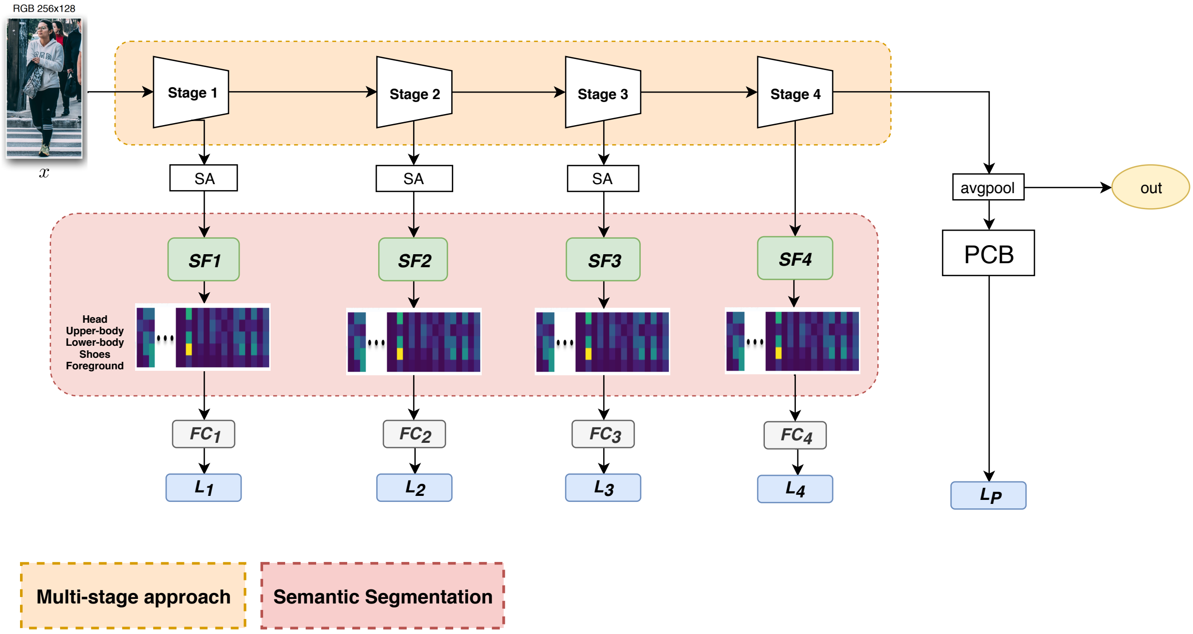

Figure 3.3: Architecture 1A: Architecture for performing semantic segmentations at multiple levels during training. SA denotes spatial attention module as proposed by *Wang et al* [45], while SF denotes using semantic segmentation to filter out information from the feature maps of the base network using masks for different body regions. We then construct a feature embedding at every stage by following the method employed by *Kalayeh et al.* [22].

The SF module spatially pools the feature maps using the weights from the semantic segmentation masks. Here we make use of the semantic masks which are down-scaled to the spatial resolution of the feature maps in a particular stage, as described in Section 3.1. This type of spatial pooling is equivalent to matrix multiplication between the feature maps and the semantic masks in the spatial domain, as described by *Kalayeh et al.* [22]. This pooling happens at each stage and produces a vector for each body part, whose length is equivalent to the number of channels in the extracted feature map. From stage 1 till stage 4 (inclusive), this corresponds to a length of 256, 512, 1024, and 2048, respectively. This part of the network is highlighted in light red color in Figure 3.3. We also made these SF modules optional, such that we can decide whether or not to use each SF module, which will be useful in our experiments in Chapter 4. If a particular stage does not make use of the SF module, it is replaced by the original training procedure, which makes use of the GAP feature of a particular stage for training.

The masking and pooling operation of feature maps using semantic masks, as mentioned in this section, is feasible because the input image undergoes multiple down-scaling operations throughout different stages of the base-network. However, the feature maps still retain their approximate semantic meaning, i.e., upper parts of the feature map will represent features that are associated with the head (if the head of the person in the input image was in that approximate location). Another approach would have been to apply these masks at the beginning, on the original input image, and independently train networks to extract features associated with each body part. However, this would result in a significantly higher computational cost, since we would have to train each one of these networks separately (a total of 5) and use them all during inference as well.

After the operations of the SF module, the features are concatenated in a similar way as described by *Kalayeh et al.* [22] before passing them through fully connected layers, i.e., we construct a fusion feature embedding by performing an element-wise max operation on the filtered feature maps of the head, upper-body, lower-body, and shoes. This intermediate feature embedding is then concatenated to the feature embeddings of the foreground and the GAP of the feature maps of that layer. However, in contrast to *Kalayeh et al.* [22], instead of only performing this operation in the final layer, we perform this operation throughout different stages of the base network.

During training, we perform a classification task using the feature embeddings at each stage, regardless of whether or not a particular stage uses the SF module. The labels of the classification task are the unique identities in the training set. Since we build upon the work of *Wang et al.* [45], we also make use of the Part-based Convolutional Baseline (PCB) module [41] in the final layer of the network. Every stage has a loss associated with it, each of which is the cross-entropy loss. The model is trained using deep supervision, similar to the approach used by *Wang et al.* [45], which back-propagates every loss term. During testing, we discard all intermediate feature maps and only make use of the *out* feature embedding (highlighted in yellow in Figure 3.3) to calculate the similarity to other images.

3.2.2. PCB Variant

In work conducted by *Bouma et al.* [1] and *Sun et al.* [41] and further explored by *Wang et al.* [45], the idea to segregate different body parts by splitting the feature maps at the final level of the base network into uniform strips has been introduced and explored. These different strips are then used for classification and consequently for training the network. *Wang et al.* [45] have made use of the Part-base Convolutional Baseline (PCB) module [41] to facilitate this idea. Since our work builds upon *Wang et al.* [45], we have also made use of this PCB, but only in the final layer of the base network, while making use of the semantic segmentation network in the lower stages.

Segregating body parts using the PCB module approach may not be as effective as using semantic segmentation since the former uses the approximate location of the different body parts. This is especially true when considering images from overhead surveillance cameras, which contain people in various different orientations. On the other hand, by chunking the feature maps in uniform strips, we can ensure that we do not mask away important information which the semantic segmentation module failed to detect. To investigate this question, we construct another network, which is a variant of *Architecture 1A*. We refer to this network as **Architecture 1B** for which a schematic representation can be seen in Figure 3.4. In this variant, we remove the semantic-segmentation module from the earlier stages and replace it with the PCB module. Feature maps at every stage are passed through a SA module, after which they undergo average pooling. They are then used in the PCB1-PCB4 modules for classification and deep supervision. Here we will experiment with different training configurations, which will be explored in Chapter 4.

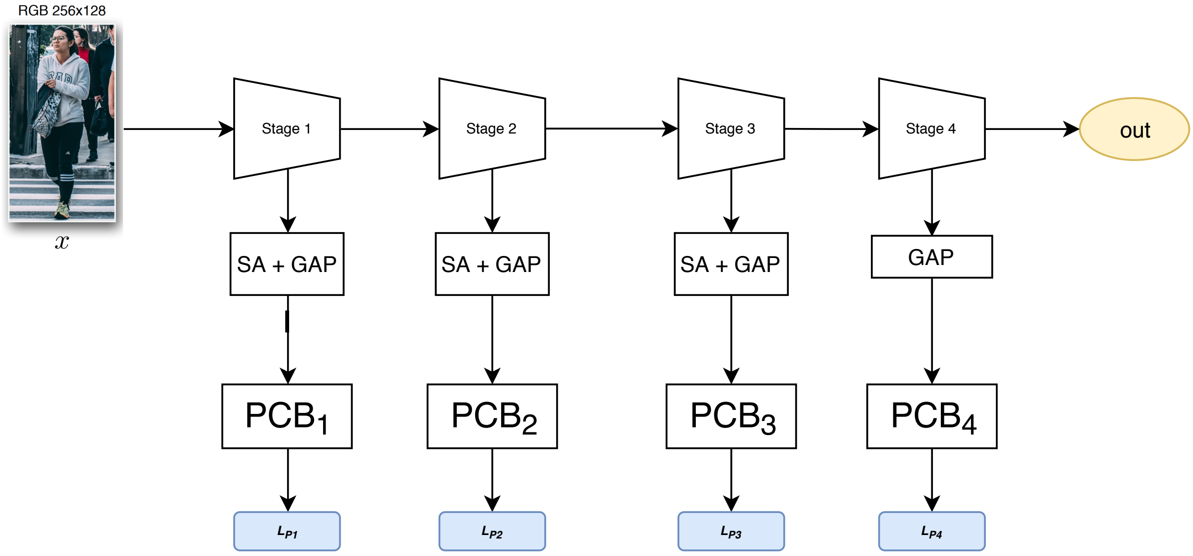

Figure 3.4: Architecture 1B: Using the Part-base Convolutional Baseline (PCB) module [41] in multiple stages of the base network as a replacement for the semantic-segmentation module.

3.3. Main Feature Extraction Network

3.3.1. Multi-branch Network

In *Architecture 1A*, we included the semantic segmentation module in multiple stages of the network. However, we argue that this is not the optimal way to use the information provided by the semantic segmentation network, i.e., it discards regional information by performing an element-wise max operation on the filtered features. Therefore, we propose an extension to *Architecture 1A*, to keep the features associated with different body parts segregated throughout the training of the re-ID network. These segregated features are optimized by training them separately to identify different persons, i.e., multiple branches. This multi-branch approach ensures that the extracted features for each body part get optimized through independent supervision, without any potential negative mutual influence due to their semantic discrepancy. We refer to this network architecture as **Architecture 2**, for which a schematic representation can be seen in Figure 3.5.

We start by using the multi-stage features extracted in *Architecture 1* to construct a feature embedding for each body part. This is done by performing a row-wise concatenation on the features after going through the SF modules, i.e., concatenating features associated with a particular body part across all stages. This produces five feature embeddings (ϕ_{head}, ϕ_{upper}, ϕ_{lower}, ϕ_{shoe}, ϕ_{fore}) , each of which uniquely represents a particular body part. We then use these features to perform multiple classification tasks for each body part. The intuition is that the network needs to learn to re-identify a person based on different body parts independently, e.g., ϕ_{head} should be able to distinguish different identities by only looking at the head. We train the network using a multi-branch approach, where each branch has its own loss, which in turn is associated with a particular body part. The loss of each feature embedding is calculated using the cross-entropy loss, which is back-propagated through the network during training. This multi-branch aspect of our network can be seen in blue color in Figure 3.5.

3.3.2. Feature Alignment during Testing

In this section, we expand on our previous ideas by not only incorporating semantic information into the model during training but also during test time. The training facilitated in our previous architecture can be seen as a way to guide the network to learn features associated with specific body parts. However, even though these networks can learn and optimize features during training, we discard them during testing and only make use of the *out* feature embedding for ranking gallery images. We argue that this approach is ineffective because we are discarding potentially important information by not considering the features that the network has learned using the semantic masks across different stages of the base network.

Several studies in the past have highlighted the problems encountered with occluded and partially

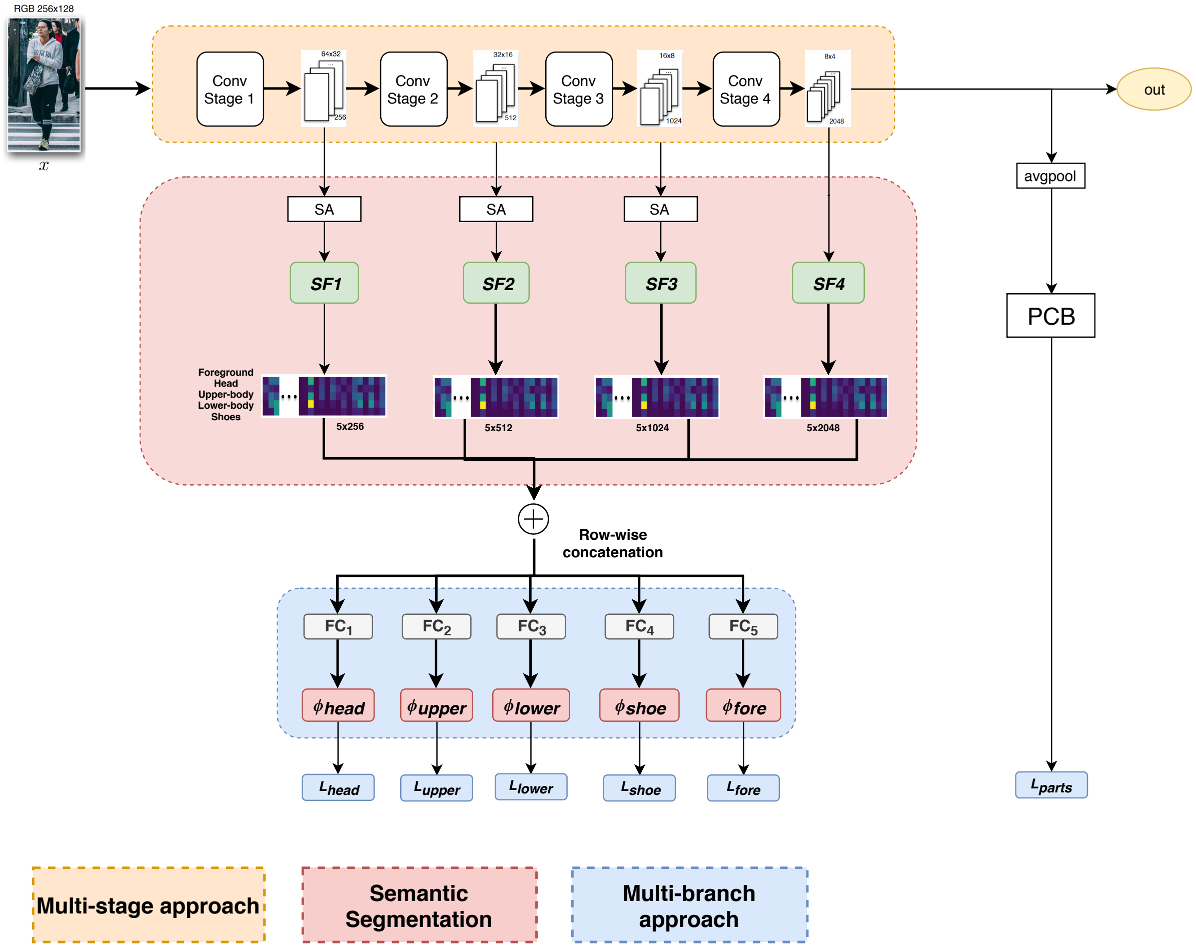

Figure 3.5: **Architecture 2:** Architecture which incorporates both multi-stage (orange), semantic-segmentation (red) and multi-branch learning (blue).

visible persons. Additionally, since we are using an external semantic segmentation network, the masks produced by this network may not always be perfect, i.e., it may not detect certain body parts or certain parts of body parts. Since we are planning to use the semantic masks during testing, these scenarios may prove to be hazardous to the performance of the model. As an example, when ranking a gallery image for which the lower-body and shoes are not visible, we do not want to assign much weight to these feature embeddings for determining the similarity between the images (and consequently influence the ranking). Similarly, if a body part is visible in a particular image, we would like to assign more weight to that body part for determining the similarity between the images. Considering this, we argue that by weighting in the visibility of body parts during the ranking process, we can obtain better re-ID performances.

To alleviate these problems, instead of computing a single similarity score using the *out* feature embedding, we compute the similarity between the different body parts independently. We do this by computing the Euclidean distance between the feature embeddings of two images for each body part. We then aggregate these similarities. The aggregation is done by weighting each similarity estimation based on the visibility of different body parts in the images. We then also consider how much influence the *out* embedding has on the final similarity estimation. The calculation of the final similarity can be seen in Equation (3.1).

$$s_{final}(i,j) = \alpha \sum_{b \in B} w_b \cdot s_b(i,j) + (1-\alpha) \cdot s_{out}(i,j) \tag{3.1}$$

where $s_{total}(i,j)$ is the final similarity estimation between image *i* and *j*. This value is dependent on the similarity estimation s_b for each body part *b* from the set of body parts *B* = *{head, upper-body, lower-body, shoes, foreground}* and the similarity estimation using the *out* feature embedding. The value of α denotes the weight we assign to the similarity estimation using the *out* feature embedding. The value of $w_b(i,j)$ denotes the importance of body part *b* when calculating the similarity between image *i* and *j*. The value of w_b is calculated using Equation (3.2).

$$w_b(i,j) = \frac{c_b(i) \cdot c_b(j)}{c_{total}(i,j)} \tag{3.2}$$

In Equation (3.2), c_b indicates the visibility of a body part, which is determined by simply counting the number of pixels assigned to it by the semantic segmentation network. We then multiply the visibility (c_b) of a certain body part in the query image *i* with the visibility of the same part in the gallery image *j*. The weight on the similarity estimation using a certain body part is then calculated by dividing by the total visibility of all the different body parts (Equation (3.3). This approach ensures that if a certain body part is barely visible in the query or gallery image, it will not be considered very important (0 in the case that a body part is not found) when computing the final similarity. Instead, its weight is redistributed across the feature embeddings of the other body parts based on their visibility in both query and gallery image.

$$c_{total}(i,j) = \sum_{b \in B} c_b(i) \cdot c_b(j) \tag{3.3}$$

4

Experiments and Results

4.1. Datasets

In the context of person re-ID, we assume that we already have the bounding boxes of persons, i.e., we do not need to train a person detector to extract images of people from surveillance footage. Consequently, benchmark datasets typically contain three sets of images: the training, query, and a gallery set, which all contain bounding box images of people. The identities found in the training set are always distinct from those in the query and gallery set. The datasets we employ during our study are Market-1501 and DukeMTMC-reID.

Market-1501 [55] is an image person re-ID dataset which is collected from 6 different non-overlapping cameras views in front of a supermarket. It consists of 1,501 identities and 32,668 bounding boxes detected using a DPM detector. Each person is captured by two to six cameras. To increase the difficulty of retrieval, the gallery set additionally contains 2,798 distractor images with just body parts or background. The **DukeMTMC-reID** [35] dataset consists of images collected from 8 high-resolution cameras containing 1,812 identities and 16,522 hand-drawn bounding boxes. Among them, 1,404 identities appear in more than two cameras, while 408 identities appear in only one camera. Examples of images contained in these datasets can be seen in Figure 4.1 and Figure 4.2 for the Market-1501 and DukeMTMC-reID dataset respectively.

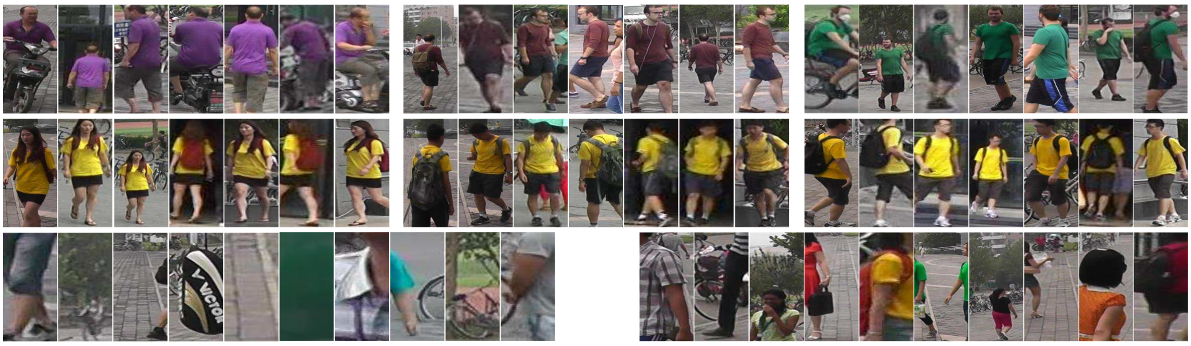

Figure 4.1: Example of images seen in the Market-1501 dataset. Image source: *Zheng et al.* [55]

4.2. Reproduction of baselines

We start our work by conducting a reproduction study on the ReSAnet network proposed by *Wang et al.* [45] and MLFN proposed by *Chang et al.* [4]. The reason for conducting this reproduction study is two-fold. First, these methods are (to the best of our knowledge) state-of-the-art in the task of person re-ID at the time of writing and will, therefore, serve as suitable baselines for comparing the performance of our method on the benchmark datasets. Secondly, both the work done by *Wang et al.* [45] and *Chang et al.* [4] make use of multi-level features available at earlier stages of the base network. Since our

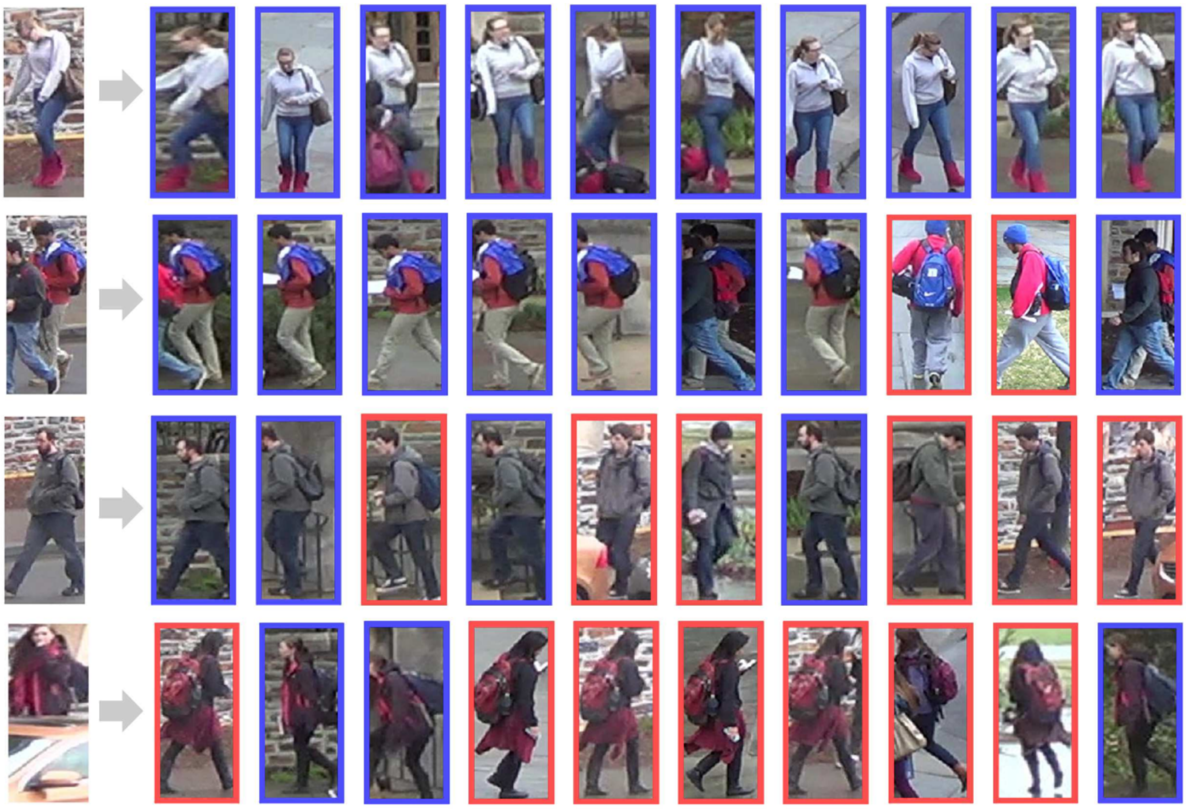

Figure 4.2: Example of images seen in the DukeMTMC-reID dataset. Image source: *Zheng et al.* [59]

proposed approach also makes use of this information, we can use one of these models as a starting point for our work.

We conduct this study by running the aforementioned networks on each benchmark dataset separately. The base network used for the reproduction study was a ResNet-50 model pre-trained on the ImageNet dataset. We have trained the complete re-ID model from scratch and tested them on the benchmark datasets with the designated testing procedure. The mAP and R1 scores of these approaches can be seen in Table 4.1. In this table, we indicate the results of the reproduction study and the results which were reported by the authors in their work between parentheses. In this table, we observe that we are able to correctly reproduce the results that were documented by *Wang et al.* [45] and *Chang et al.* [4]. We can also observe that the ReSAnet architecture [45] achieves better results compared to MLFN architecture [4], which follows a similar approach. Therefore, we shall use the ReSAnet architecture [45] as our baseline and also as the starting point for our proposed method.

Datasets	**Duke**		**Market1501**	
Method name	mAP	R1	mAP	R1
MLFN [4]	63.2 (62.8)	81.1 (81)	74.3 (74.3)	90.1 (90.0)
ReSAnet [45]	84.8 (85.9)	87.7 (89)	90.1 (91.7)	93.8 (94.7)

Table 4.1: Reproduction study of past research on several benchmark datasets. In parenthesis are the results documented by the authors in the original study.

4.3. Experiments

In this section, we explore the possibility of not only incorporating semantic information in the final feature maps of the base network but also in the feature maps of the earlier stages. We argue that these "semantically complemented features" can be used to train the network to extract more discriminative feature embeddings. The final goal of these experiments is to investigate whether semantic information can be used to complement the base network to address the feature alignment problem better. We start by providing semantic information on various body parts of the person across multiple stages of the base network during training. Later, we experiment by also incorporating this semantic information during testing, which allows the network to compare different feature embeddings, which are associated with different body parts of the person, thereby addressing the feature alignment problem.

4.3.1. Semantic Segmentation at Multiple Stages

The purpose of our first experiment is two-fold: (1) investigate whether incorporating semantic information in lower stages of the base network during training leads to better performance and (2) investigate which of these stages benefits the most from semantic information during training. We argue that the use of semantic segmentation in the final stage of the base network will lead to degraded performance due to the low spatial resolution. We also argue that the use of semantic segmentation in lower stages will lead to improved results due to the more fine-grained information available in these layers. We refer to this experiment as **Experiment 1**.

For this experiment we make use of *Architecture 1A* as described in Section 3.2.1. We start the experiment by using semantic masks only in the final stage of the base network, hereafter referred to as experiment 1.1. Here we "turn-off" the SF1-SF3 modules, which indicate the use of semantic segmentation in those stages. This experiment can be thought of as a naive combination of the method suggested by *Wang et al.* [45] and *Kalayeh et al.* [22]. We then proceed by progressively incorporating semantic masks to lower levels of the network (experiment 1.1 - 1.4). Additionally, starting at the final layer, we remove the usage of semantic masks on these layers and progressively make our way down to the lowest layer of the base network (experiment 1.4 - experiment 1.7). We document the results of the aforementioned experiments in Table 4.2. The SF1-SF4 columns in this table correspond to the SF modules depicted in the Figure 3.3.

	Configuration				DUKEMTMC-reID		Market-1501	
Technique	SF1	SF2	SF3	SF4	mAP (%)	R1 (%)	mAP (%)	R1 (%)
ReSAnet	no	no	no	no	70.6	84.6	77.8	92
experiment1.1	no	no	no	yes	70.3	84	78	92.7
experiment1.2	no	no	yes	yes	69.8	83.5	78.2	92.6
experiment1.3	no	yes	yes	yes	69.9	84.4	77.9	91.7
experiment1.4	yes	yes	yes	yes	69.2	83	76.9	92.2
experiment1.5	yes	yes	yes	no	68.7	82.4	78.1	92.8
experiment1.6	yes	yes	no	no	67.7	81.3	78.2	92.4
experiment1.7	yes	no	no	no	66.7	80.7	76.9	91.9

Table 4.2: Performance results for executing multi-level experiments. SF denotes that we have used semantic segmentation to filter out information from the feature maps at a certain level of the base network using masks from different regions from the body.

From Table 4.2, we observe that by applying semantic segmentation at the lowest level of the base network (experiment 1.7), the performance decreases considerably with respect to the baseline. Furthermore, by adding semantic information to even higher layers, we can see a slight increase in performance (experiment 1.4 - 1.6); however, it is still considerably less than the baseline model. A different effect is observed when progressively adding semantic information to the lower levels, starting from the last level. In this case, the results are inconclusive as they are fairly similar compared to the performance of the baseline model.

From these experiments, we can see that including semantic information on the earliest stage has a detrimental effect on the performance of the model. A possible explanation for this is that the feature maps constructed at this level still contain too much noise since it is still in the early stages of the base network. Therefore, this disadvantage may outweigh the benefit that these feature maps contain more fine-grained information, thereby leading to the observed degradation. On the other hand, features at later stages of the base network are more abstract and contain less noise, making the SF module less likely to include noise in the filtered feature maps. This also explains the similar results between the baseline model and the experiments 1.1 - 1.3.

4.3.2. Multi-branch Experiments – Segregating Body Parts

The purpose of this experiment to investigate the effectiveness of keeping feature maps associated with different body parts segregated during training. Unlike experiment 1, we argue that by taking into account the semantic discrepancy between different regions of a feature map, we can improve the performance of the network. Additionally, we want to investigate whether the semantic-segmentation module can be used to complement or replace the learning concept proposed by the PCB. The PCB module works by splitting the final feature map in six uniform strips with the idea to capture the approximate location of each body part. Since the idea behind the PCB module is similar to our approach, we argue that using semantic segmentation will lead to better performance due to its pixel-level accuracy. This experiment is referred to as **Experiment 2**.

To validate our hypotheses, we make use of *Architecture 2* as described in Section 3.3.1. However, in experiment 1, we concluded that using feature maps at the final stage of the network and the last three stages produced the best results, i.e., experiments 1.1 and 1.3, respectively. Therefore, in this experiment, we slightly adapt *Architecture 2* to only make use of semantic masks in the stages as they are described in experiment 1.1 and experiment 1.3. The stages which do not use the semantic-segmentation module make use of the traditional learning procedure. We also experiment whether this approach is more effective than the PCB by removing the loss terms of the latter. The results of Experiment 2 are documented in Table 4.3.

				Duke		Market1501	
Method name	Multi-branch	Base	PCB	mAP	R1	mAP	R1
ReSAnet	no	-	-	70.6	84.6	77.8	92
experiment1.1	no	-	-	70.3	84	78	92.7
experiment1.3	no	-	-	69.9	84.4	77.9	91.7
experiment2.1	yes	1.1	yes	69.6	83	78.2	92.5
experiment2.2	yes	1.1	no	66.1	80.5	71.3	87
experiment2.3	yes	1.3	yes	69.3	82.3	78.3	92.5
experiment2.4	yes	1.3	no	67	80.8	70.1	86.7

Table 4.3: Performance results for segregating body parts during training of the network. Base refers to the underlying architecture used during a particular experiment, while PCB denotes whether we make use of the parts based classifier.

In this table, we indicate in which stages we include semantic masks during training by referring to the architectures used in Experiment 1 in the column named “Base”. Additionally, we indicate whether or not we train using the PCB module in the “PCB” column. From the results, we can see that segregating feature maps from different body parts in the architecture of experiment 1.1, does not result in a better performance (experiment 2.1). Likewise, keeping features segregated across the last three levels of the base networks also does not increase the performance, as we had hypothesized (experiment 2.3). Finally, we observe that even though the purpose of the current experiment is to improve upon and replace the PCB network, removing the PCB network has a detrimental effect on the performance of the network when considering both architectures of our previous experiment (experiment 2.2 and 2.4). Therefore, from this experiment, we can conclude that keeping different body parts segregated during training does not help in learning more discriminative feature embeddings and that the function of the PCB network cannot be improved/replaced by the semantic segmentation network.

4.3.3. Feature Alignment – Last Stage

In this experiment, we investigate whether we can improve the performance of our model by not only using “semantically complemented features” in the lower stages of the network during training but also use these learned features during testing. We also investigate whether we can further increase the performance by making use of the semantic-segmentation module in assigning importance per body part during testing. Finally, we investigate whether we can complement or replace the *out* feature embedding by using the features learned for each different body part during test time. We argue that the features constructed using the semantic-segmentation module are superior due to their pixel-level accuracy and should thereby lead to better performance. We also hypothesize that by using weights for each body part, we can achieve better performance, since the model can now account for scenarios where certain body parts are not visible. This experiment is referred to as **Experiment 3**.

For this experiment we make use of *Architecture 2* as described in Section 3.3.1. Here we make use of the variant we used in experiment 2.1, which incorporates the semantic-segmentation module only in the last stage of the network. We also make use of our new feature embedding during testing, as described in Section 3.3.2. To investigate the effectiveness of this new feature embedding, we first use the equal distribution of weights (not considering the visibility) across the similarities estimated using different body parts. The results for using equal weights can be seen in Table 4.4, under experiment 3.1 and experiment 3.2. Here we make a distinction between not making use and making use of the *out* feature embedding, respectively. We repeat the same experiment, but this time using weights based on the visibility of different parts in experiment 3.3 and experiment 3.4 (without and with *out* feature embedding respectively).

	Multi-branch				Duke		Market	
Method name	train	test	weight	out	mAP	R1	mAP	R1
ReSAnet	no	no	-	yes	70.6	84.6	77.8	92
experiment2.1	yes	no	-	yes	70.3	83.9	78.2	92.5
experiment3.1	yes	yes	equal	no	56	75.5	63.8	84
experiment3.2	yes	yes	equal	yes	68.7	82.7	78.1	91.8
experiment3.3	yes	yes	weighted	no	57.4	76.4	68.7	87.4
experiment3.4	yes	yes	weighted	yes	70.4	84.1	77.9	91.9

Table 4.4: Using architecture of experiment 2.1 as base, we show the performances of using weighted averages on similarity estimation of different body parts. Additionally we also document the performances when using and not using the *out* feature embedding for similarity estimation.

In Table 4.5 the “train” column indicates whether we use multi-branch training as used in Experiment 2, while the “test” column indicates whether the learned features for each body part are also used during testing. The “out” column indicates whether we make use of the *out* feature embedding during testing. From these results, we observe a similar outcome as in experiment 2 when removing PCB during training i.e., we observe a significant performance degradation when we do not consider the out feature embedding during testing. We observe a slight performance degradation when comparing experiment 2.1 with experiment 3.1. However, we do see a slight performance improvement with respect to experiment 2.1 when we include weights based on the visibility of body parts while still using the *out* feature embedding during testing (experiment 3.4). From these observations, we can conclude that the *out* feature embedding is still essential during testing and cannot be replaced by the feature embedding trained using semantic segmentation. We can also conclude that weighting similarity scores based on the visibility of the body parts does give a slight performance improvement when testing with and without the *out* feature embedding. Unfortunately, the architecture and testing procedure proposed in this experiment does not seem to outperform our baseline.

4.3.4. Feature Alignment – Multiple Stages

In this experiment, we investigate the same research questions as the *Experiment 3*. However, in this experiment, we will make use of the architecture used for experiment 2.3. We argue that using features from lower stages of the network during testing will improve the performance of the model since it has access to more fine-grained information in these stages. This experiment is referred to as **Experiment 4**.

For this experiment we again make use of *Architecture 2*, as described in Section 3.3.1. The setup of this experiment is very similar to the one used in experiment 3. However, instead of using the architecture from experiment 2.1, we use the architecture from experiment 2.3, which trains using features across the last three stages of the base network. Similar to experiment 3, we will make a distinction between weighting the similarity estimation equally versus weighting them based on the visibility of body parts in both the query and gallery images. We also make a distinction between using and not using the out feature embedding during testing. The results of this experiment can be seen in

Table 4.5.

	Multi-branch				Duke		Market	
Method name	train	test	weight	out	mAP	R1	mAP	R1
ReSAnet	no	no	-	yes	70.6	84.6	77.8	92
experiment2.3	yes	no	-	yes	69.3	82.3	78.3	92.5
experiment4.1	yes	yes	equal	no	54.7	74.2	62.3	83.3
experiment4.2	yes	yes	equal	yes	70.6	84	78	91.9
experiment4.3	yes	yes	weighted	no	57.3	71.1	68.7	87.1
experiment4.4	yes	yes	weighted	yes	70.4	84.2	77.9	92

Table 4.5: Using architecture of experiment 2.3 as base, we show the performances of using weighted averages on similarity estimation of different body parts. Additionally we also document the performances when using and not using the *out* feature embedding for similarity estimation.

The columns in this table have the same meaning as in Experiment 3, i.e., "train" and "test" columns indicate whether we use multi-branch training and testing respectively. The "out" column indicates whether we also use the *out* feature embedding for calculating a similarity score. From the results in this table, we observe that the removal of the *out* feature embedding during testing again results in a degradation of performance. We also observe a slight improvement when using weights for the different similarity scores with respect to equal weight distribution. Unfortunately, the conclusion for this experiment is that the usage of features from earlier stages during testing does not result in better performance when comparing to the baseline.

4.3.5. Semantic Segmentation Revisited

In this experiment, we investigate whether the semantic-segmentation module has any influence during the training of the network. Since all of our previously suggested methods perform worse than our baseline, we argue that the use of semantic segmentation is not beneficial for the training of the network and might even have a detrimental effect. This experiment is referred to as **Experiment 5**.

To validate our hypothesis, we again make use of *Architecture 1A* as described in Section 3.2.1. However, instead of creating semantic masks for each body part using the semantic-segmentation module, we use maps filled with ones that essentially allow all information to pass through. This results in the regular GAP instead of GAP with filtering using semantic segmentation. Consequently, each level will now have a much larger feature representation, which is used for classification. Compared to the baseline model, this is a 5x increase in feature representation size. This setup allows us to study the influence of increasing the size of the feature embedding, without including semantic information. The results for this experiment can be seen in Table 4.6.

		Duke		Market-1501	
Method name	Filter	mAP	R1	mAP	R1
ReSAnet	-	70.6	84.6	77.8	92
experiment1.3	semantic segmentation	69.9	84.4	77.9	91.7
experiment5.1	ones	69.3	83.8	77.9	92.5

Table 4.6: Comparison between approaches with/without filtering using semantic information. Experiment 1.3 used the semantic-segmentation module in the lower stages of the network. We use the same model for experiment5.1, but replace the semantic masks with masks filled with ones (i.e. let through all the information.

From these results, we can see slight performance degradation in the Duke dataset when removing the semantic masks, while no effects can be observed in the Market dataset. It may be the case that the baseline model is already capable of identifying different body parts of the person in an image and does not need semantic information for learning better features. We, therefore, argue that ineffective training is caused by the fact that some of the semantic segmentation masks contain noise or cannot segment certain body parts correctly. This leads to unnecessary information being filtered out (or let through), which decreases the effectiveness of training.

Therefore, from this experiment, we can conclude that the use of semantic segmentation has no positive effects on the performance of the re-ID model. We can also conclude that, compared to the baseline, the feature dimension also does not affect the performance of the model.

4.3.6. PCB in Multiple Stages

Our final experiment explores whether the PCB module can be used in multiple stages of the base network to increase the performance of the model. We also explore different strategies for incorporating the PCB in multiple stages of the base network during training. This experiment is referred to as **Experiment 6**.

We have made use of *Architecture 1B* as described in Section 3.2.2. We have experimented with several configurations of this approach, of which the results can be seen in Table 4.7. In this table, experiment 6.1 indicates that each level has six feature representations from the PCB, which is used to train the network. In experiment 6.2, we increment the length of the feature representations of the PCB with each additional stage. These lengths correspond to 256, 512, 1024, 2048, for stage 1 till stage 4, respectively. Finally, experiment 6.3 performs a weighted average for each part of the PCB across all layers before performing the classification. The weights are set on the feature maps of each stage and are trainable parameters which are optimized during the training phase.

				Duke		Market1501	
Method name	Stage 1-3	Stage 4	Method	mAP	R1	mAP	R1
ReSAnet	-	PCB	-	70.6	84.6	77.8	92
experiment1.4	SS	SS + PCB	-	69.2	83	76.9	92.2
experiment6.1	PCB	PCB	Separated	65.7	80.8	73.7	90.2
experiment6.2	PCB	PCB	Separated (incremental)	65.6	81.2	72.7	89.8
experiment6.3	PCB	PCB	Weighted	67.2	81.9	73	89.6

Table 4.7: Performances of replacing the semantic segmentation networks with the PCB modules on each level of the base network. The first two columns indicate the differences between the different architectures in terms of where the PCB is places (ReSAnet vs experiment 1.4 vs experiment6)

From the results in this experiment, we can see that integrating the PCB across all stages of the base network degrades the performance of the model with respect to ReSAnet. We can also observe that having different sizes of feature representations across the different levels also does not improve the performance. Finally, by performing a weighted average of the feature representations of the PCB increases the performance slightly, but is still worse than ReSAnet. In light of these results, we can conclude that adding the PCB to multiple stages of the base network degrades the performance of the model with respect to both the ReSAnet baseline and our previous experiment 1.4. The reason for this degradation may be that the noise present in the early stages of the network has more negative impact on the performance when using the PCB module. The performances are less adverse when using the semantic segmentation module, because the latter can limit the amount of noise better than the PCB module.

4.4. Discussion

In light of the results from our experiments, we found that incorporating semantic segmentation in multiple stages of the base network does not improve the performance of the model beyond the state-of-the-art. Incorporating semantic segmentation in the first stage, in particular, proved to be detrimental to the performance. This conclusion also holds when replacing the semantic-segmentation module in each stage with PCB-modules, which uses a similar idea. This degradation may originate from the fact that the lower stages of the network contain more noise than the higher stages which are detrimental to the overall performance. The degradation is more prominent when using the PCB module instead of the semantic-segmentation network and can be accounted by the fact that the PCB module allows more information to pass than the semantic-segmentation module.

We also found that segregating features associated with different body parts and optimizing them using a multi-branch approach during training did not improve the performance of the network. Furthermore, we also concluded that the multi-branch approach could not be used to replace the training facilitated by the PCB-module in the final layer, despite their similar role in the training process and the seeming advantage the multi-branch approach has due to its use of semantic segmentation. We found that using features associated with different body parts during testing does not help the model improve with respect to the baseline. This conclusion also holds for weighting feature embeddings based on the importance of a particular body part. Finally, in our last experiment, we found that the semantic segmentation network does not provide any benefits to the re-ID model during training. The reason for this may be that the re-ID network is already capable of extracting semantic information and does not need the additional information provided by the semantic segmentation network. Another argument may be that the benefits gained by good quality semantic masks are outweighed by the lower quality semantic masks, which introduce much more noise during training.

5

Person re-ID on MatchNMingle

The **MatchNMingle** dataset created by *Cabrera et al.* [2] was originally intended for the analysis of social interactions in the wild. However, this dataset poses an interesting challenge in the context of person re-ID. As opposed to current person re-ID benchmark datasets, which mainly contain images taken from a slightly elevated viewing angle, this dataset contains person images taken from down-facing overhead cameras. Examples of images from this dataset can be seen in Figure 5.1. This viewpoint causes a large portion of the lower-body to be occluded from the cameras in certain frames, thus providing less information for re-ID purposes. Given the viewing angle, people appear to be in different orientations depending on where they are located in the frame. These points emphasize the feature misalignment problem. Despite the widespread use of this type of surveillance in practice, researchers have currently made no effort to investigate these scenarios, nor is there any dataset available to facilitate this work.

In this chapter, we contribute by introducing a new dataset that covers a context in person re-ID which is not yet considered in current person re-ID research. We detail how we create a re-ID dataset from the *MatchNMingle* dataset. We also conduct several experiments and document the preliminary results.

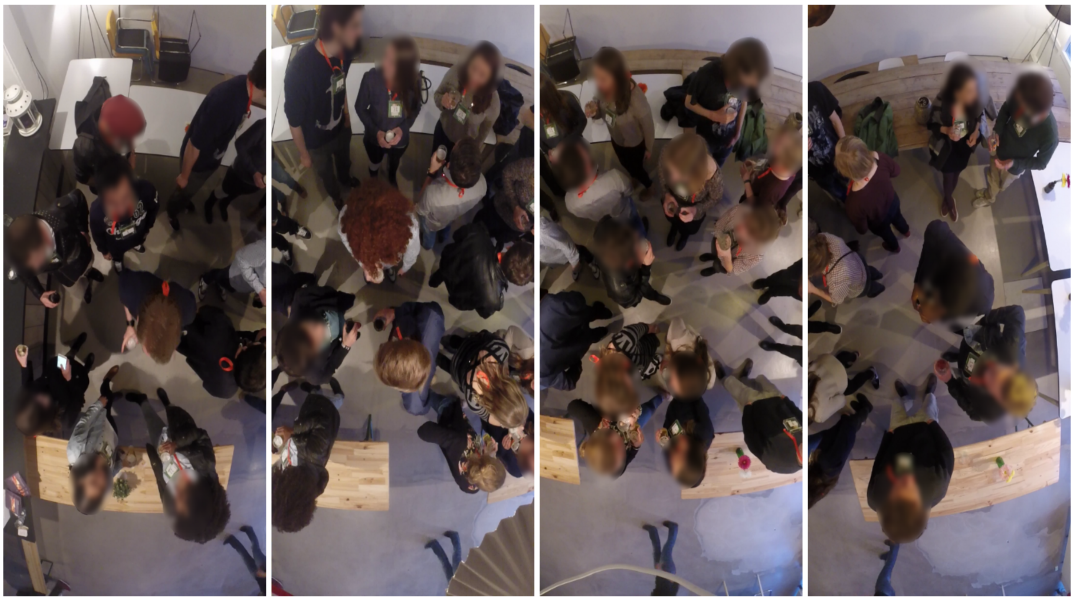

Figure 5.1: Examples of images found in the MatchNMingle dataset. Image source: *Cabrera et al.* [2]

5.1. Creating MatchNMingle-reID

The MatchNMingle dataset is originally intended for studying social interaction between groups of people. The recordings in MatchNMingle are made for a period of 30 minutes using 3 cameras over 3 days, resulting in a 1.5 hours recording for each camera. Each day includes a different set of participants. Transforming this data into a re-ID dataset is challenging in itself since there are “only” 92 participants in the dataset, much less than the most widely used person re-ID datasets. The number of cameras and duration is also less than the current widely used benchmark datasets. Furthermore, several people do not move a lot across the duration of the social experiment, thereby limiting the diversity of views obtained from a certain subject. Nonetheless, this dataset remains interesting since it allows us to perform person re-ID from a whole different viewing angle. Therefore, we have taken a few steps to limit the impact of these limitations on the quality of the resulting re-ID dataset.

Most of the current re-ID datasets (DukeMTMC-reID, Market-1501, MARS), have a 50/50 train-test split of the identities at their disposal. These datasets also work with cross-camera search mode, which ensures that the images from the query dataset and the gallery set are from different cameras. We have taken a similar approach when constructing the re-ID version of this dataset. In the first version of the new re-ID dataset, we chose to use the 50/50 train-test split as proposed in previous datasets. However, we do not guarantee cross-camera search mode i.e., images from a particular identity in the query and gallery set may be taken under the same camera. This choice is based on the fact that most identities do not tend to move a lot across cameras, and while some do, the limited number of cameras would limit the number of query images we can sample from the dataset.

Considering the aforementioned points, we first start the creation of the re-ID dataset by randomly selecting participants for the train and testing sets. For the training set, we then randomly sample 100 images for each participant from the entire sequence (under all cameras) when they are in view. For the query set, we consider the entire video sequence under all cameras where a particular identity is in view and create seven equally sized splits. We then use the first frame of each split and the last frame of the final split as a query image, resulting in 8 query images per identity. From the remaining images, we randomly sample 10 images per split for the gallery set, resulting in 70 gallery images per identity. To ensure that the query images do not look identical to the gallery images, we exclude frames during sampling, which occur within 30 seconds from the query image. Figure 5.2 schematically depicts how this sampling strategy is implemented. Finally, we perform a visual inspection to remove indiscernible identities from all three sets.

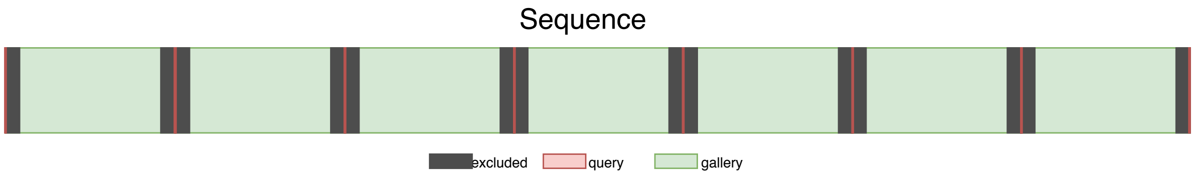

Figure 5.2: Sampling strategy for MatchNMingle-reID query and gallery set. The full rectangle denotes the entire 30-minute video sequence of a particular person. Red indicate the query image, and green indicates the times where gallery images are sampled, and black indicates frames which are excluded.

5.2. Base Networks vs ReSAnet

Since the MatchNMingle-reID dataset consists of several new challenges unseen in previous re-ID datasets, we will begin by testing the performances of stripped-down backbone networks such as the ResNet50 and XceptionNet. Since most existing state-of-the-art methods make use of modules that are highly tailored to the widely used re-ID datasets, this experiment may give us insights into how we can improve these networks to tackle the problems encountered in this new re-ID dataset, while also retaining good performance on the other datasets. Since the type of images is much different from traditional person re-ID datasets, we argue that these stripped-down networks will have comparable, if not better, results to the current state-of-the-art network.

We conduct this experiment by training and testing these networks on the MatchNMingle-reID dataset. We do not make use of our model, which we presented in Chapter 3, since the semantic masks generated for the images of MatchNMingle-reID are extremely poor and would not give us any meaningful insights. In order to compare to the current state-of-the-art, we also evaluate the performance of ReSAnet on MatchNMingle-reID. We present these results in Table 5.1.

	MatchNMingle-reID	
Technique	mAP (%)	R1 (%)
ResNet50	9.9	9.5
XceptionNet	17.7	27.4
ReSAnet	12.6	16.2

Table 5.1: Base network and ReSAnet performance on MatchNMingle-reIDv1 dataset

In this table, we can see that, although the performance of ResNet50 comes close to the performance of ReSAnet, XceptionNet seems to outperform the state-of-the-art method. These results also confirm that the current state-of-the-art is highly optimized for re-ID tasks in traditional datasets, which is expected given its use of the PCB network as a sub-module. With this experiment, we can also conclude that we cannot entirely rely on the insights obtained from techniques used in previous research when working with the MatchNMingle-reID dataset.

5.3. Rotation Correction

In the previous experiment, we have seen that ReSAnet does not outperform a standard XceptionNet and that this is likely because ReSAnet is optimized for previous re-ID datasets that have upright frontal views of identities. In MatchNMingle-reID, the identities are not always upright and mainly have only their upper-bodies visible to the camera. Since ReSAnet makes use of the PCB, which implicitly assumes approximate locations of body parts, it fails to learn anything meaningful since these images appear in different orientations. Therefore, in our next experiment, we propose to perform a rotation correction to the images of the MatchNMingle-reID dataset, which would allow ReSAnet to learn more meaningful features.

The bounding boxes in the MatchNMingle dataset are manually annotated, which gives us information on the position of a particular identity. Since we want the persons in the images to stand upright, we calculate an angle and rotate the image along this angle. An example of such rotation can be seen in Figure 5.3. This is a valid assumption to have in practice since a network detecting bounding boxes around participants will also have information about their location. Using these rotation corrected images, we again run our experiment as in the previous section i.e., comparing performances of standard feature extractor networks with the state-of-the-art ReSAnet. The results of this experiment can be seen in Table 5.2.

	MatchNMingle-reID	
Technique	mAP (%)	R1 (%)
ResNet50	9.4	10.1
XceptionNet	19.1	27.9
ReSAnet	21.9	30.2

Table 5.2: Base network and ReSAnet performance on MatchNMingle-reIDv1 dataset when images are rotation correct

In this table, we observe that the performances of the standard feature extractors drop slightly, while the performance of the state-of-the-art increases slightly and thereby outperforming the standard feature extractors. These results are again to be expected since ReSAnet now has the opportunity to effectively make use of the PCB module, which assumes upright persons in the dataset. However, even though the performance of ReSAnet has increased using this simple trick, the performance improvement is relatively low, and overall performance is still much lower than other person re-ID datasets. Nonetheless, from this experiment we can conclude that although the rotational correction technique improves the performance of ReSAnet slightly, it is clear that there are still many other challenges left to solve.

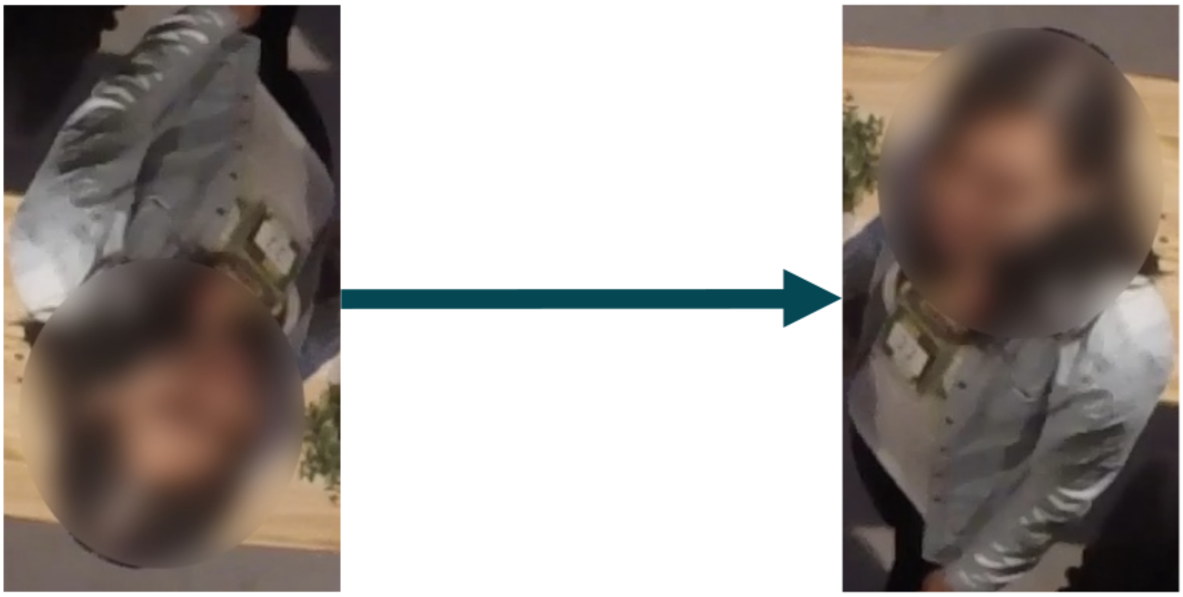

Figure 5.3: Example of performing rotation correction on a given image in the MatchNMingle-reID dataset.

5.4. Random Rotation

In the previous section, we observed that, although rotation correction helped ReSAnet obtain a better performance, the models still perform relatively bad on the MatchNMingle-reID dataset. This can be attributed to the fact that the models do get better at understanding images where the head is always at the upper part of the image, but are still capable of learning features for the new viewing angle. Therefore, we argue that it might be better to train the network to work with all the challenges found in the dataset, but providing it with more examples through data augmentation. The augmentation that we employ in this experiment is a random rotation around a random angle on the training set. The performances of the models can be seen in Table 5.3. From this table, we observe that data augmentation has helped to increase the performance as compared to the non-augmented dataset. XceptionNet also appear to improve slightly compared to its performance using the rotational correction technique. However, from this experiment we can also conclude that data augmentation through random rotation alone is not sufficient to tackle the problems encountered in this dataset.

	MatchNMingle-reID	
Technique	mAP (%)	R1 (%)
ResNet50	10.4	14.5
XceptionNet	21.1	33.0
ReSAnet	21.0	27.9

Table 5.3: Base networks and ReSAnet performance on MatchNMingle-reIDv1 dataset using random rotation augmentation.

5.5. Dataset Validation

In the previous sections, we observed that the performances of the models are relatively low compared to more traditional datasets. Therefore, in this section, we want to investigate what the cause is for this discrepancy in performance. We start by training and testing on the train set of the MatchNMingle-reID dataset. The purpose of this experiment is to verify if a model is capable of learning features required to perform re-ID on MatchNMingle-reID. If this is the case, the performance should be nearly perfect, since the model will be tested with identities it has seen during training. Therefore, we train an XceptionNet using the full training set, after which we divide it into a query and gallery set. After testing on these sets, we obtained the results which are presented in Table 5.4. In this table, we can see that the results are nearly perfect, which suggests that the network is capable of extracting relevant information by training on the available data. We can also visually observe that the model is capable of generating discriminative feature embeddings for identities in the training set from the t-SNE plot shown in Figure 5.4.

	MatchNMingle-reID	
Technique	mAP (%)	R1 (%)
ResNet50	97.5	100

Table 5.4: Testing on the training set of the MatchNMingle dataset.

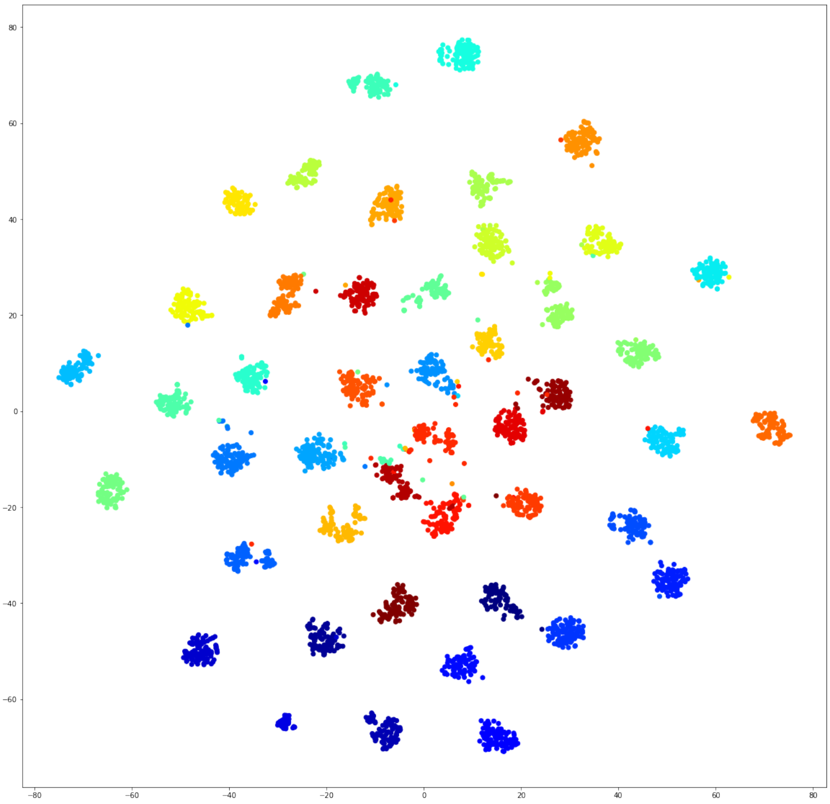

Figure 5.4: t-SNE plot of identities in the training set.

We further investigate the performance discrepancy by considering the number of identities that are available in the training set of MatchNMingle-reID. A possible explanation for this discrepancy can be explained by the fact that there are fewer identities in the training set of MatchNMingle-reID when compared to e.g., DukeMTMC-reID dataset (46 vs. 751 training identities). In order to verify this assumption, we conduct an experiment which involves training on various number of identities of the training set and testing on the test set. In this experiment, we start with a small number of training identities that are randomly sampled from the training set, after which we test on the testing set. This process is repeated by gradually adding more identities to the training set until we train with all the identities available in the complete training set. The performance of XceptionNet model as a function of the number of unique identities is presented in Figure 5.5. In this figure, we can see that for a small number of identities (50) in the training set, XceptionNet reveals a performance close to that which is achieved on the MatchNMingle-reID dataset. Therefore, from this experiment, we can conclude that the low performance on the test set originates from the small number of training identities in the MatchNMingle-reID dataset.

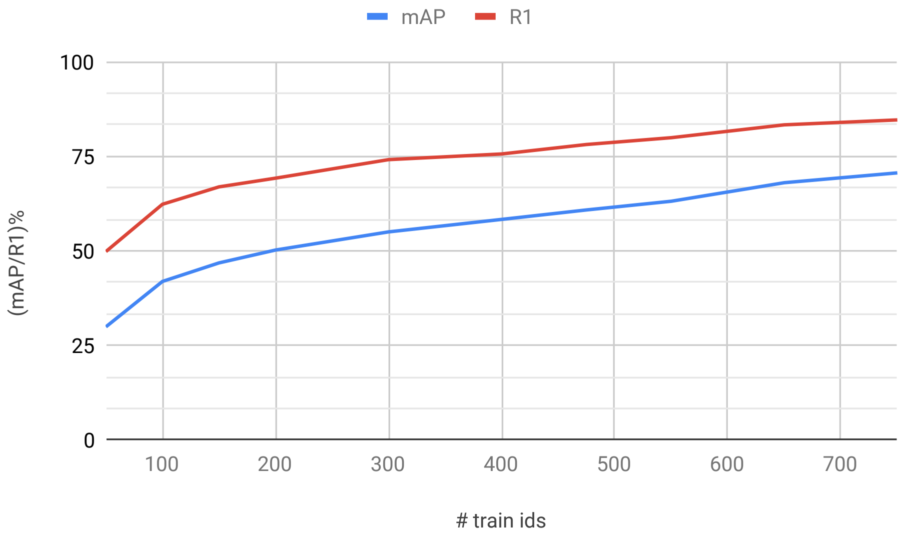

Figure 5.5: mAP and R1 performance plotted against the number of identities in the training set.

5.6. Discussions

In our initial experiments, we found that a basic XceptionNet feature extractor can outperform the current state-of-the-art ReSAnet on the MatchNMingle-reID dataset. However, by augmenting the dataset, we found that the performance of ReSAnet and XceptionNet were comparable. This may indicate that the initial poor performance of ReSAnet may be due to the lack of identities and pose variety in the dataset. This speculation is also backed by our last experiment, which shows the performance curve of ReSAnet on DukeMTMC-reID when varying the number of identities in the dataset. However, results obtained from our second experiment suggests that ReSAnet is optimized to perform well in cases where people are always standing upright and may not be the best choice when looking at images of people taken from overhead cameras. Therefore, due to the limited number of identities in the dataset, it still unclear whether any of these speculations hold when looking at datasets with a larger number of identities and more variety. Nonetheless, in this chapter, we can conclude that the MatchNMingle-reID dataset does not have enough identities to make a fair comparison between different datasets. However, with our contribution, we hope to encourage the person re-ID community to start investigating person re-ID by creating larger datasets or better models, which are more generalizable in a similar context with minimal number of training samples.

6 Conclusion

6.1. Conclusions

In our work, we have investigated the effectiveness of using semantic information to complement a standard feature extractor to better address the feature alignment problem. We use a multi-stage approach that is used to extract more fine-grained features encoded in earlier stages of the feature extractor network. Our multi-branch approach is used to optimize the features extracted for different body parts of a person. We also introduce a novel testing procedure which takes into account the importance of the feature embedding for each body part before ranking the gallery set. With the results obtained from our experiments, we empirically show that semantic segmentation may not be the correct choice when trying to address the feature alignment problem. This can be concluded from our experiments, which suggest that the use of a semantic-segmentation module during training does not give any additional performance improvements when compared to our baseline. There are also no noticeable performance improvements when we weight the features of distinct body parts based on their importance.

In addition to a new network architecture, we also introduced a new person re-ID dataset named MatchNMingle-reID. This dataset covers a new context for person re-ID, which currently has not received any attention from the re-ID community. Preliminary results indicate that the number of identities in the MatchNMingle-reID dataset is relatively small compared to other traditional datasets to make any fair comparisons and draw conclusions. Despite this limitation, our experiments show that there are still many interesting problems left to solve in these types of datasets, which may inspire new and more generalizable models for person re-ID in the future.

6.2. Future Work

While our work has mainly focused on addressing the feature misalignment problem, we argue that a model that can truly generalize well across different datasets will also need to solve several other problems that are not discussed in our work. These problems include a change in illumination between frames, background biases, and domain adaptations. Past research (Appendix A.2) have also shown that there is a lot to be gained in terms of performance with re-ranking approaches. Therefore, a truly robust model will likely require a combination of multiple different approaches, including data augmentation, model improvements, and re-ranking. Finally, any such optimized model can be reused to perform person re-ID when there is more information available in the form of videos. Past research, discussed in Appendix A.3, has documented the challenges and benefits of performing person re-ID based on videos.

With our work, we hope to lay the groundwork for the re-ID community to build models that generalize well to datasets with significantly different viewpoints. There are several points which can be improved on and also some challenging problems to consider. Due to the limited number of identities in the MatchNMingle dataset, we chose not to guarantee cross camera search as this would limit the number of query images in the test set. However, in order for the dataset to be representative of challenges faced in practice, we highly recommend future work to focus efforts on creating more massive datasets with similar viewpoints, which are representative of real-world scenarios. This includes increasing the number of cameras, changing the context in which the videos are recorded, and increasing the total recording time. This can, in turn, be used to create a re-ID dataset that can be used to test models using cross camera search, which is more representative of real-world scenarios. Instead of creating larger datasets, it could also be interesting to increase the generalizability e.g., with transfer learning, to perform domain adaptation with a minimal number of training samples. Another interesting problem arises in creating a dataset that contains surveillance footage taken from overhead cameras and traditional viewpoints. This would capture not only an individual in different orientations (arbitrary vs. upright), but also different environments (indoor vs. outdoor). We argue that such a dataset will test the limits of current state-of-the-art approaches and inspire even more generalizable models for person re-ID.

Bibliography

[1] Henri Bouma, Sander Borsboom, Richard JM den Hollander, Sander H Landsmeer, and Marcel Worring. Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense XI*, volume 8359, page 83590Q. International Society for Optics and Photonics, 2012.

[2] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[4] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, pages 2109–2118, 2018.

[5] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018.

[7] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BVMC*, 2011.

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[9] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *NIPS-workshop 2017*, 2017.

[10] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *European conference on computer vision*, pages 330–345. Springer, 2014.

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994–1003, 2018.

[12] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535. IEEE, 2006.

[13] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[16] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 41–47, 2007.

[17] Yiluan Guo and Ngai-Man Cheung. Efficient and deep person re-identification using multi-level similarity. *CVPR 2018*, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[21] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

[22] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[25] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[26] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[27] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[28] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018.

[29] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*, volume 1, pages 192–196. ACM, 2016.

[30] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[31] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[32] Andy J Ma, Pong C Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *Proceedings of the IEEE international conference on computer vision*, pages 3567–3574, 2013.

[33] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, number 5, 2010.

[34] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. *arXiv preprint arXiv:1803.10859*, 2018.

[35] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[36] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 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*, pages 420–429, 2018.

[37] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6886–6895, 2018.

[38] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. *CVPR*, pages 5363–5372, 2018.

[39] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[40] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017.

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

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[43] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5794–5803, 2018.

[44] Arthur van Rooijen, Henri Bouma, and Fons Verbeek. Fast and accurate person re-identification with Xception conv-net and C2F. In *IberoAmerican Congress on Pattern Recognition CIARP*, 2018.

[45] Haoran Wang, Yue Fan, Zexin Wang, Licheng Jiao, and Bernt Schiele. Parameter-free spatial attention network for person re-identification. *arXiv preprint arXiv:1811.12150*, 2018.

[46] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, pages 2275 – 2284, 2018.

[47] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8042–8051, 2018.

[48] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018.

[49] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[50] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[51] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[52] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018.

[53] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6781–6789, 2018.

[54] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 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*, pages 1077–1085, 2017.

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[56] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 2019.

[57] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2012.

[58] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[59] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[60] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3652–3661. IEEE, 2017.

[61] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[62] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[63] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.

A

Additional Related Work

A.1. Data Augmentation

One form of data augmentation is to perform domain adaptation through style transfer using Generative Adversarial Networks (GANs), in particular the CycleGAN architecture [9]. In this task, given a source dataset S and a target dataset T, the goal is to alter the images in dataset S such that they still contain the same identity, but have the style (such as illumination, background and image noise) of dataset T. Following this line of reasoning, the work carried out by *Deng et al.* [11] uses a combination of CycleGAN and a Siamese Network, called Similarity Preserving Generative Adversarial Network (SPGAN), to generate samples which not only possess the style of the target domain but also preserve their underlying ID information. The CycleGAN is used to translate an annotated dataset S from the source domain to target domain in an unsupervised manner such that a new labeled training dataset G(S) is created on the target domain. In addition to the adversarial and cycle consistency loss, the authors proposed a target domain identity constraint (another loss term which works as regularization) which empirically shows not produce unrealistic images and preserves color composition. The generated images (from source to target) are then passed through the Siamese Network which computes the contrastive loss. These 4 losses are then back-propagated to train the complete network. Once the source image has been translated to the target image, it is fed through the Siamese Network for feature learning and re-ID.

The work proposed by *Liu et al.* [30] aims to enrich the training data with a wide range of human pose variations. While existing approaches using GANs only consider whether a produced image looks realistic or not (unrelated to re-ID) and therefore do not preserve identity information, this work proposes a pose transferable person re-ID framework which utilizes pose transferred sample augmentations (i.e., with ID supervision) to enhance re-ID model training. This is performed by estimating pose with a pre-trained model as suggested by *Cao et al.* [3] and transferring them onto a static human image using ideas from the conditional GAN (cGAN), but trained to also maintain the identity of the generated image with new pose. The GAN is also given an extra *Guider* module to guide the GAN to adapt to the re-ID problem i.e. boost its discriminative power using cross-entropy or triplet loss. This GAN is trained on the MARS dataset (to gain the source poses) and used to generate new pose images on new target person re-ID datasets for robust training.

A.2. Reranking

Zhong et al. [60] proposed the k-reciprocal encoding method to re-rank the re-ID results. The hypothesis is that if a gallery image is similar to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match. Specifically, given an image, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector. These features are then used for re-ranking under the Jaccard distance. This is done by looking at the expanded k-reciprocal nearest neighbor and creating a weighed vector for each image in the gallery. The final distance is computed as the combination of the original distance and the weighed Jaccard distance.

Saquib et al. [36] proposed an improvement to the previous re-ranking approach with a new unsupervised and automatic re-ranking framework called the Expanded Cross Neighborhood Distance based Re-Ranking which does not require to compute new rank lists for each image pair (e.g., based on reciprocal neighbors), but instead looks at the accumulated distance of the the immediate two-level neighbors of each image with the other image and is much less computationally expensive than methods proposed using Jaccard Distance [60]. They also propose an approach that incorporates both the fine and coarse pose information of the person to learn a discriminative embedding.

A.3. Video

Video data may contain much richer information about pedestrian appearance and also convey motion clues that implicitly reflect persons body layouts and can be beneficial for robust person re-ID. However not all frames may contain useful information, some may be redundant, others may have noise. To better distill relevant information from the videos, recent works introduced attention mechanisms to video-based person re-identification.

Chen et al. [5] focuses on the problem that a single vector may not be sufficient to embed the visual variation in an entire video sequence. The authors solve this by dividing the sequence into multiple video snippets, embed each snippet separately, and perform snippet-similarity estimation using DNN with co-attentive mechanism on snippet embedding and aggregation. With this scheme, the intra-sample visual variation could be minimized while the diverse appearance and temporal information of the sequence are maintained. Probe and gallery sequences are broken down to snippets, then the intersnippet similarity is calculated using DNN with co-attentive embedding and sorted on similarity value of which the top ranked similarities are averaged, producing a similarity score between sequence p and q. The similarity between 2 snippets are computed by first extracting features from each frame using a CNN. The query feature is learned from the whole probe snippet by an LSTM network thus is aware of the overall appearance and motion of the snippet. The key pairs and the query feature are used are used to generate the attention weights for the value features. The combination of the attention weights and value features of the snippet are then used to create the snippet embedding.

Typical person re-ID methods usually describe each pedestrian with a single feature vector and match them in a task-specific metric space, which is usually not sufficient to overcome visual ambiguity. *Si et al.* [38] proposed an end-to-end trainable framework, called Dual ATtention Matching network (DuATM), to learn context-aware feature sequences and perform attentive sequence comparison simultaneously. The core component of our DuATM framework is a dual attention mechanism, in which both intra-sequence and inter-sequence attention strategies are used for feature refinement and feature-pair alignment, respectively. The dual attention block computes the feature aware map (given a reference feature) of an image and uses that to compute the attention weight map for both the intra-class refinement and interclass alignment. These vectors are computed in both direction (S1 to S2 and S2 to S1) such that comparisons in both directions can be made when defining the distance of the holistic sequence. Thus, detailed visual cues contained in the intermediate feature sequences can be automatically exploited and properly compared. The proposed DuATM network is a Siamese network using triplet loss assisted with a decorrelation and cross-entropy loss. Decorrelation loss is used reduce overfitting and also give a compact representation for feature sequence. Cross-entropy loss with data-augmentation is used to get more robust and informative features.

The work done by *Wu et al.* [50] proposes to exploit unlabeled tracklets by gradually but steadily improving the discriminative capability of the CNN feature representation via step-wise learning. The proposed method first trains a CNN model on a one-shot labeled tracklet, by trying to classify the identity correctly (which helps create the initial feature embedding). A technique called EUG (Exploit Unknown Gradually), then iteratively improves the feature representations formed by the CNN by two steps, the label estimation step and the model update step. In the first step, EUG generates the pseudo labels for unlabeled tracklets, and selects some of pseudo-labeled tracklets for training according to the prediction reliability. The tracklets are labeled by averaging over every extracted feature embedding of each frame and calculating it's distances to the labeled feature representation. The selected subset is continuously enlarged during iterations according to a sampling strategy, which looks at the reliability of the psuedo-labels. In the second step, EUG re-trains the CNN model on both the labeled data and the sampled pseudo-labeled subset in order to improve the feature representation for the next iteration. The authors employ a progressive sampling method to increase the number of the selected pseudo-labeled candidates step by step.

Zhang et al. [53] proposed an interpretable reinforcement learning based approach to person re-ID. An agent is trained to verify a pair of images at each time. The agent could choose to output the result (same or different) or request another pair of images to verify (unsure). The model implicitly learns the difficulty of image pairs, and postpone the decision when the model does not accumulate enough evidence. Furthermore, by adjusting the reward for unsure action, a trade off can be made between speed and accuracy. The model makes decisions, by first extracting the features of an image using either Inception or AlexaNet. The states are determined by the extracted feature vector and the weighed averages of the difference between the historical frames. The rewards are determined by predicting a match correctly or postponing the final decision.

Most of the work done in video-based re-ID, realize the problem that a single feature vector is not sufficient to encode the complex movements of a person throughout a video sequence due to pose variations and occlusions. The most successful techniques have therefore proposed a selective feature extraction approach, where the features are extracted with the aid of attention networks or multi-branch approach where each branch extracts independent features with the aid regularization. From this section it is clear that a video based person re-ID model can benefit significantly by taking into account that occlusions need to be accounted for and extracted features have to be aligned when incorporating temporal information.