Person Re-identification in Images and Videos

by

Tanzila Rahman

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

Thesis advisor

Dr. Yang Wang

Author

Tanzila Rahman

Person Re-identification in Images and Videos

Abstract

Person re-identification is a challenging task of matching a query person across multiple person's images or videos captured from different camera views. Recently, deep learning based approaches have showed promising performance on this task. In this thesis, initially we propose an image based person re-identification approach with Spatial Transformer Networks. Most previous deep learning based approaches use whole image features to compute the similarity between images. This is not very intuitive since not all the regions in an image contain information about the person identity. Hence, we introduce an end-to-end Siamese convolutional neural network that firstly localizes discriminative salient image regions and then computes the similarity based on these image regions. Furthermore, we propose an efficient attention based model for person re-identifying from videos. Our method generates an attention score for each frame based on frame-level features. The attention scores of all frames in a video are used to produce a weighted feature vector for the input video which is refined iteratively for re-identifying persons from videos. Extensive experiments on different datasets show that the proposed models provide an effective way of re-identifying person from images as well as videos.

Contents

List of Figures

List of Tables

Acknowledgments

At first all, I offer heartfelt gratitude to Almighty God who gave me the ability to complete my whole thesis work without any breaking.

I am very grateful to get the opportunity of work under the supervision of Dr. Yang Wang. I would like to express my sincere appreciation and gratitude to him for his indispensable guidance, unparallel stimulating influence and continuous encouragement throughout the progress of my research work. He is a great academician and mentor who lead me in the field of computer vision and deep learning. Without his thoughtful reviews and guidance I would hardly finish this thesis work. Thank you Prof. Yang Wang, I will always be grateful for all that I have learned from you.

I would also like to thank my committee members, Dr. Carson Kai-Sang Leung and Dr. Jun Cai, for their valuable suggestions and constructive feedback in perfecting my thesis. Thanks are also to all of support staff in the Department of Computer Science for their helping hand.

I also take this opportunity to express my gratefulness to - The University of Manitoba, Faculty of Graduate Studies and The Government of Manitoba for their continuous financial support.

I would like to extend my gratitude to my wonderful lab mates for their time, support and suggestions and for being so nice, kind and helpful.

Last but not the least I would like to express our hearty complement to my parents and family members for their unconditional love and encouragement to come at this stage and pursue my dream.

This thesis is dedicated to my parents
for their unconditional love and endless support.

Publications

Some of the ideas, materials and figures in this thesis have appeared previously in the following publications and submitted manuscripts:

1. **Tanzila Rahman**, Mrigank Rochan and Yang Wang. Person Re-Identification by Localizing Discriminative Regions. *The 28th British Machine Vision Conference (BMVC)*, London, United Kingdom, September 2017.

2. **Tanzila Rahman**, Mrigank Rochan and Yang Wang. Video-based Person Re-identification Using Refined Attention Networks. *11th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2018)*, IIIT Hyderabad, India, 2018. (Under review)

Chapter 1

Introduction

Person re-identification is the problem of recognizing a specific person in a system of non-overlapping camera views. It is an important problem in many real-world applications, such as video surveillance, human computer interaction, police investigation and so on. It is a very challenging problem due to the complex variations in viewpoints, poses, lighting, illuminations, blurring effects, and image resolutions. The intra-person variations can even be larger than inter-person variations in this task [1]. Backgrounds and occlusions also create challenges in person re-identification.

In this thesis, we consider two person re-identification problems. One is image based person re-identification and another one is video based person re-identification. The goal of image based person re-identification is to identify a specific person in an input image (known as the probe image) from a set of images (known as gallery set) captured by non-overlapping and different cameras. Sometimes small objects or regions convey important information about the person identity in an image. Humans can recognize person identity based on these salient regions. For example, in Fig.1.1,

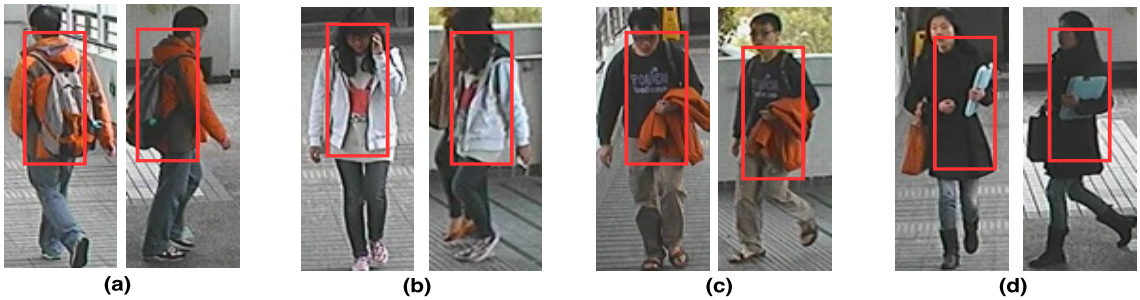

Figure 1.1: Some examples of pedestrian images for image based person re-identification from CUHK01 train dataset. Each pair represents the same person from different camera views. The bounding box on each image shows the discriminative region localized by our proposed approach.

person (a) carries a backpack, person (b) wears a white jacket, person (c) holds an orange colored jacket in his hand and person (d) holds a file in her hand. These distinctive regions can be used to identify one person from others. Usually, if an object is salient in one camera view, it remains salient in another camera view too [2] even though there are variations in view points. In addition to salient objects, body parts as well as clothing can also be considered as informative region for identifying persons. Although salient regions in an image play a vital role in person re-identification for humans, most existing approaches in person re-identification do not capture this information. Most of the existing approaches [3; 4; 5; 6; 7] compute the similarity between two images based on whole image features.

In this thesis, we propose a new person re-identification technique by explicitly localizing salient regions. In particular, we use Spatial Transformer Network (STN) [8] to localize the discriminative regions in the input images. Our multichannel CNN model then computes the similarity of the input images based on these discriminative regions in conjunction with whole image features.

Most of the earlier work (e.g. [9; 10; 11; 12; 13; 14; 15; 16]) focuses on the image based re-identification. Recently, video-based person re-identification is receiving increasing attention (e.g. [17; 18; 19; 20; 21; 22; 23; 24]). Compared with static images, video-based person re-identification is a more natural setting for practical applications such as video surveillance.

In this thesis, we also consider the problem of video-based person re-identification. Given a video containing a person, the goal is to identify the same person from other videos possibly captured under different cameras. A common strategy for person re-identification is to formulate it as a metric learning problem. Given the query video and a candidate video, the goal is to develop algorithms to compute the distance between these two videos. If the distance is small, it means the two videos likely contain the same person. See Fig. 1.2 for an illustration.

Previous work (e.g. [21; 23]) has made the observation that not all frames in a video are informative. For example, if the person is occluded in a frame, ideally we would like the feature representation of the video to ignore this frame and focus on other "useful" frames. A natural way of solving this problem is to use the attention models [25; 26; 27] that have been popular in visual recognition recently. In [21; 23], RNN is used to model the temporal information of the frames and generate the attention score for each frame for person re-identification.

In this thesis, we propose a new attention model for video-based re-identification. Compared with previous works [21; 23], our model has several novelties. First, instead of using RNN, we directly produce the attention score of each frame based on the image feature of this frame. Since the attention score of each frame is calculated

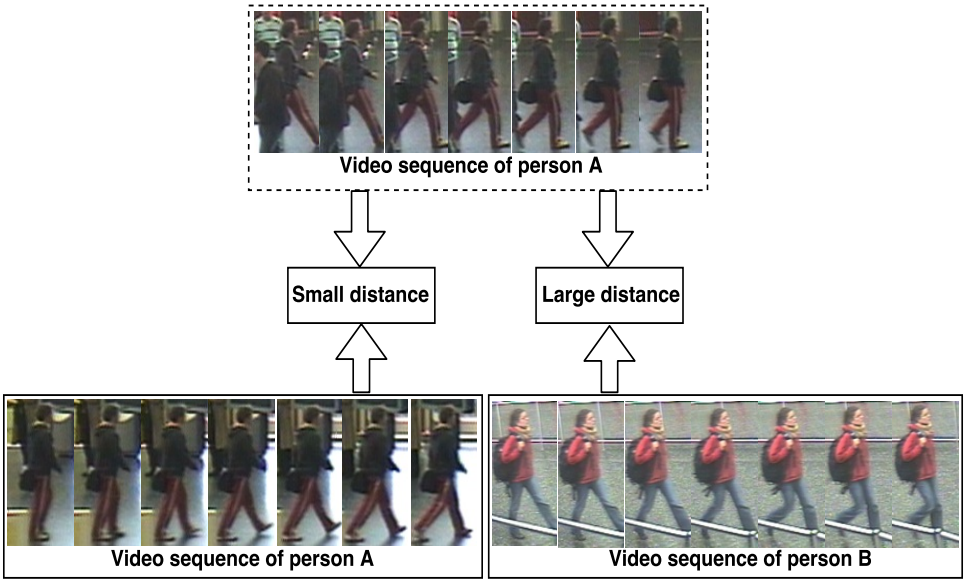

Figure 1.2: Illustration of the video-based person re-identification problem. In this case, our goal is to identify person A from two video sequences in the second row. If two videos contain the same person, we would like the distance between them to be small. Otherwise, we would like the distance to be large. Some frames in a video sequence may be affected by occlusions and are not informative about the person's identity. In this thesis, we use an attention model to focus on informative frames for re-identification.

based on the frame, the computation of attention scores over all frames can be easily made parallel and take full advantage of the GPU hardware. Second, the work in [21; 23] only calculates the attention scores once. Here, we introduce a new method to refine the attention scores based on the whole video features. We show that this attention refinement can improve the performance of our model.

1.1 Contributions

Our contributions include:

- We integrate attention-based STN in the image based person re-identification

framework. This allows our model to focus on discriminative regions in the input images when computing their similarity. Moreover, we integrate global image features with the discriminative regions to produce final feature representation for person re-identification. To the best of our knowledge, this is the first CNN-based architecture that performs person re-identification by localizing discriminative image regions. Our model can be trained end-to-end and it does not require supervision or any prior knowledge about the discriminative regions.

- A new attention mechanism for video-based person re-identification. Unlike previous work (e.g. [23]) that uses RNN to generate the attentions, our model directly generates attentions based on frame-based features. As a consequence, the computation of the attentions is much simpler and can be easily parallelized. In contrast, RNN has to process frames in a sequential order, so the computation cannot be made parallel. Despite of its simplicity, our model outperforms the more sophisticated RNN-based attention mechanism in [23]. We also introduce an iterative refinement process to further improve the attentions. This allows the model to refine the attention scores over time. We show that this attention refinement improves the performance of the final model. In addition, we also study the effect of iterative refinement on the performance.

1.2 Thesis Organization

The remainder of the thesis is organized as follows: First, in Chapter 2, we review the related works and the literature. In this chapter, we briefly describe previous

approaches and techniques used for both image and video re-identification. In Chapter 3, we describe image based person re-identification by localizing discriminative regions. We show that discriminative regions in the input images can play vital role for re-identification. Hence, we introduce a Siamese convolutional neural network to localize discriminative salient regions and later use that regions in conjunction with the whole image to re-identify person. In Chapter 4, we develop deep learning architectures to address video re-identification. Here, we propose an attention based architecture that generate attention scores for frame level features. Unlike most existing deep learning methods that use global or spatial representation, our approach focuses on attention scores. We also propose an iterative refinement approach and show that this attention refinement improves the performance of the final model. Finally, we conclude this thesis in Chapter 5.

Chapter 2

Related Work

Previous work in person re-identification falls into two broad categories: image-based re-identification and video-based re-identification. In this Chapter, we briefly describe some previous approaches that are related to re-identifying person from both images and videos.

2.1 Image-based Person Re-identification

There has been extensive work on person re-identification from static images. Early work in this area uses hand-crafted feature representations [28; 29; 30; 31; 32]. Most of these methods involve extracting feature representations that are invariant to viewpoint changes, then learning a distance metric to measure the similarity of two images.

Deep learning approaches, in particularly deep convolutional neural networks (CNNs), have achieved tremendous successes in various visual recognition tasks [33].

In many areas of computer vision, CNNs have replaced hand-engineering feature representations with features learned end-to-end from data. Recently, CNNs have been used for image-based person re-identification [9; 10; 11; 12; 13; 14; 15; 16]. These methods use deep network architecture such as Siamese network [34] to map images to feature vectors. These feature vectors can then be used for re-identification. Hence, previous work on person re-identification can be classified into two broad groups: non-deep learning methods and deep learning methods.

Non-Deep Learning Methods: Most of the person re-identification methods consist of two components: (1) a method to extract features from the input images, and (2) a way of computing a similarity metric to decide whether the images belong to same person or not. Much of the previous research focuses on either improving feature extraction method [35; 36; 37; 38], or robust similarity metric learning [39; 40; 41; 42; 36], or their combination [3; 43; 44; 45]. Although these approaches are promising, their performance is limited due to the heavy reliance on handcrafted features. In contrast, our approach is based on deep learning which simultaneously learns the feature representation and a similarity metric to optimize the performance.

Deep Learning Methods: In recent years, deep neural networks have significantly improved the state-of-the-art in many computer vision tasks such as image classification and object detection. There are a few previous works that use deep learning for person re-identification problem in the literature. Our work is mostly related to the work by Yi et al. [46], Li et al. [47], Ahmed et al. [5], and Subramaniam et al. [7]. Yi et al. [46] propose a Siamese convolutional network for re-identification. Their network takes a pair of images as its input to which three stages of convolution

are performed followed by a fully-connected operation that outputs a vector for each input image. Lastly, cosine similarity function is used to compare the two output vectors. Li et al. [47] use a two-input network architecture that firstly performs a set of convolutions to the inputs and then multiplies the convolution feature maps at different horizontal offsets. This is followed by a max-out grouping which filters out the highest response from horizontal strips to which another convolution and max pooling operation is done. Finally, the output is used to compute the similarity. Ahmed et al. [5] introduce a deep architecture that contains two new layers: cross-input neighborhood layer and patch summary layer. Cross-input neighborhood layer is used to learn the relationship between feature maps of two input images. Patch summary layer is responsible for summarizing the neighborhood maps by analysing the differences in each 5×5 block, which are then used to measure the similarity of two input images. Our image based re-identification model is motivated by recent work in [7] which extends the work of [5]. The work in [7] uses a fused network that performs inexact matching through a novel layer called Normalized X-Corr whose output assists the subsequent layers in making decision on whether the two input images are similar or not. The main difference between these previous approaches and ours is that, instead of using only whole image feature maps to compare the two input person images similarity, we firstly localize discriminative regions in the images and then forward their feature maps in addition to the global images to subsequent layers for similarity computation. Our work is driven by the intuition that the input images contain a lot of background pixels which are irrelevant for person re-identification.

Our work is also related to the recent work on localizing and ranking visual at-

tributes given a pairwise image comparison [48]. This work uses STN to localize the image regions that are relevant for the visual attribute. Similar to [48], we also incorporate STN to localize discriminative regions in images that are relevant for person re-identification.

2.2 Video-based Person Re-identification

Although the performance of image-based person re-identification has increased significantly, this is not a very realistic setting for practical applications. To address the limitation of image-based re-identification, a lot of recent work has began to explore video-based re-identification [17; 18; 19; 20; 21; 22; 23; 24] since it is closer to real-world application settings. Compared with static images, videos contain temporal information that is potentially distinctive for differentiating a person's identity. Some prior work has explored ways of incorporating temporal information in deep convolutional neural network for re-identification. For example, McLaughlin et al. [19] use CNN on each frame in a video and incorporate a recurrent layer on the CNN features. Temporal pooling is then used to combine frame-level features into a single video-level feature vector for re-identification.

One work is also related to a line of research on incorporating attention mechanism in deep neural networks. The attention mechanism allows the neural networks to focus on part of the input and ignore the irrelevant information. It has been successfully used in many applications, including machine translation [25], image captioning [27], visual question answering [26], etc. In video-based re-identification, the attention mechanism has also been explored [21; 23]. The intuition is that only a small portion

of the video contains informative information for re-identification. So the attention mechanism can be used to help the model focus on the informative part of the video.

The work in [23] is the closest to ours. It uses an RNN to generate temporal attentions over frames, so that the model can focus on the most discriminative frames in a video for re-identification. In this thesis, we uses temporal attentions over frames as well. But instead of using RNN-based models to generate attentions [23], we directly calculate the attention scores based on frame-based features. This makes the model much simpler and the computation of attention scores can be easily parallelized over frames. We also propose an attention refinement mechanism to iteratively refine the attention scores. We demonstrate that this attention refinement improves the performance of the final model.

Chapter 3

Person Re-Identification by Localizing Discriminative Regions

We formulate person re-identification as a binary classification problem given two input images. Our proposed model learns a function f that maps an image pair (I_1, I_2) to a score that indicates how likely these two images correspond to the same person. During training, our network takes an image pair (I_1, I_2) and a binary label L indicating whether the images are similar or not. During testing, the input is an image pair (I_{test1}, I_{test2}) and the network uses the learned function f with parameters w to predict the similarity score $f(I_{test1}, I_{test2})$ between the image pair.

Figure 3.1 shows the overall architecture of our model. Our model is based on the Siamese network [49]. It has two Siamese networks whereas each network contains two branches with shared parameters. There are two main components in the network: (a) *Spatial Transformer Network* (SN) and (b) *Fused Network* (FN). The STN is used to learn to localize the discriminative region in an image and generate a feature

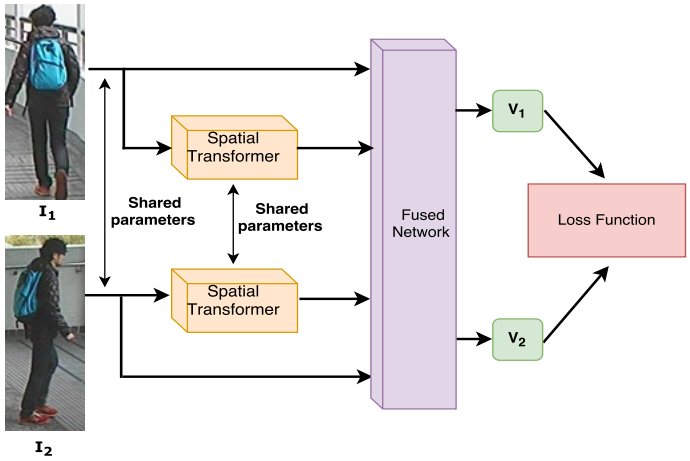

Figure 3.1: Overall architecture of our network. It takes two person images (I_1, I_2) as its input. Each image is forwarded to two Siamese-CNN architecture whereas one contains a Spatial Transformer Network (STN) with a Fused Network (FN) and another contains only fused network. The model finally produces two outputs/scores (v_1, v_2) indicating similarity strength of two input person images which is later fed to a loss function to update the parameters of the network.

representation based on this region. The FN is used to combine the features of discriminative regions in both input images and output a similarity score.

3.1 Spatial Transformer Network

Previous work on person re-identification typically compares the similarity of two images based on features extracted from the entire image. We believe this is not optimal, since an image usually contains a lot of pixels (e.g. background pixels) that are irrelevant for person re-identification. Humans usually differentiate between a pair of images by focusing on certain distinct regions/parts of the person in the image (see Fig. 1.1). In our work, we develop a model that has the same capability. In our model architecture, we incorporate STN for localizing discriminative regions that are

relevant for person re-identification. STN is a fully-differentiable module that can learn spatial transformations, such as scaling, rotation and translation without any additional supervision.

We incorporate STN in our network so that it can focus on discriminative regions which would be used for subsequent parts of the network. The output of STN will simplify the task of Fused network (FN) as it can be optimized efficiently over the localized discriminative regions for a given pair of images.

As outlined in [8], there are three main components in STN (see top of Fig. 3.2): i) Localization network, which takes the input image and produces the transformation parameters θ; ii) Grid generator, which generates a sampling grid using the transformation parameters. The sampling grid is a set of points where the input feature map should be sampled to produce the transformed output; and iii) Sampler, which uses a bilinear interpolation kernel to produce the output image. In this work, we use STN that has three transformation parameters $\theta = [s, t_x, t_y]$, where s, t_x and t_y represent isotropic scaling, horizontal and vertical translation respectively. This transformation parameters are constrained for attention [8], and the point transformation is

$$\begin{pmatrix} {x_i}^{in} \\ {y_i}^{in} \end{pmatrix} = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \begin{pmatrix} {x_i}^{out} \\ {y_i}^{out} \\ 1 \end{pmatrix} \tag{3.1}$$

Here, ${x_i}^{in}$ and ${y_i}^{in}$ represent coordinates of the input image, whereas ${x_i}^{out}$ and ${y_i}^{out}$ represent output image coordinates at the i-th index. The localization network within the STN can take any form of convolutional network or fully-connected network, but finally it should include a regression layer that generates transformation parameters

θ [8]. In this thesis, we follow their localization layer architecture which uses STN for digit localization in images. A convolutional layer with 20 filters of size 5×5 and two fully-connected layers are added towards the end. The first fully-connected layer takes 6120 values as its input and produces 20 output values, whereas the second one takes 20 input values and produces 2 transformation parameters (t_x, t_y) as output. Here the network is not learning scaling parameter (s) as we fix it to 0.5. These parameters are used to generate the transformed output image patch through the sampling mechanism. Figure 3.3 shows the STN's localization behavior during training.

3.2 Fused Network

The input images and the output of STN are fed to the Fused Network (FN) [7]. In our model, we use two fused networks separately. One of them takes a pair of image patches as its input whereas another one considers a pair of whole input images. Finally these two fused network outputs the similarity score indicating whether two image belong to the same person or not. The fused network is also a Siamese network where each branch contains two stage of convolutions (with shared parameters) and pooling layers. These convolution layers take input image of size $60 \times 160 \times 3$ and generate 25 feature maps of dimension 12×37 which is fed to the normalized correlation and the cross-input neighborhood layers. Given two feature maps, the normalized correlation layer [7] computes the correlation between every pair of 5×5 patch matrices. For given matrices X and Y, the Normalized Correlation is defined as [7]:

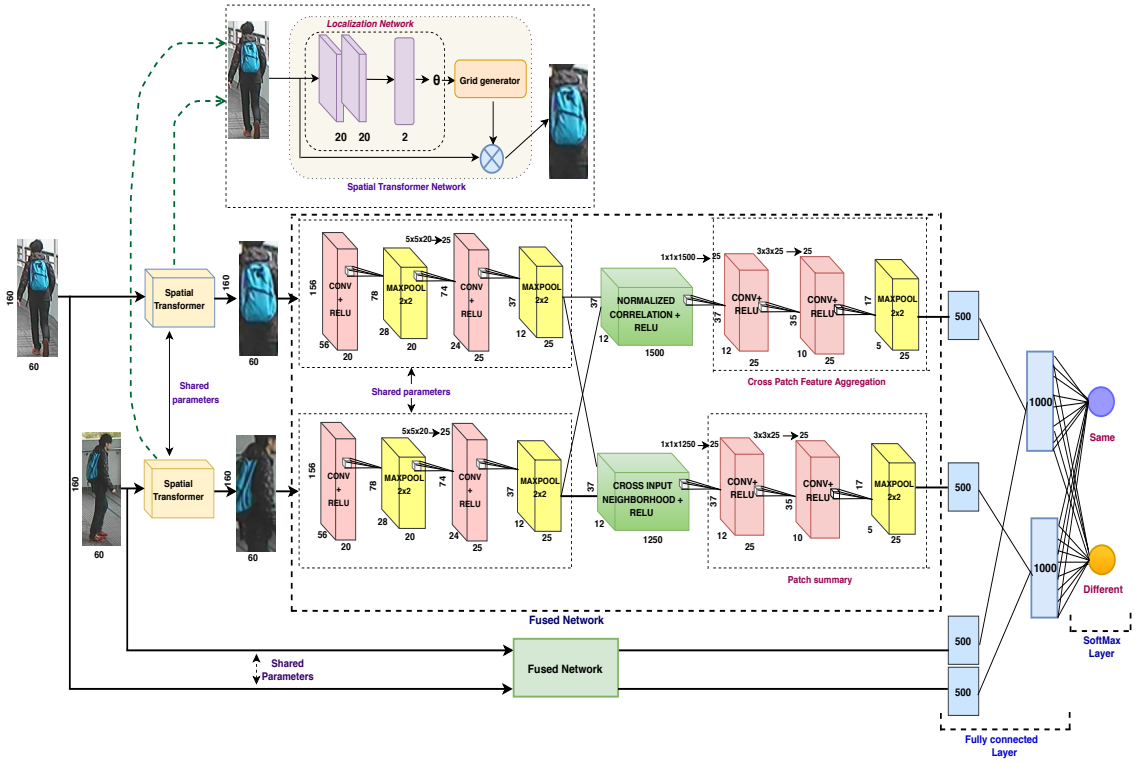

Figure 3.2: Detailed architecture of our proposed network. The network takes a pair of image as input. Each image goes through the spatial transformer network (STN), which localizes the discriminative image region. The output of STN is fed to Fused network which generates two linear layers with 500 output values as features of the discriminative region. At the same time, the input images go through another fused network which also produces two linear layers of 500 output values as global image features. The features from the localized regions and the global images are concatenated and finally used to compute the similarity score of the two input images.

$$normCorr(X, Y) = \frac{\Sigma_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{(N-1).\sigma_X.\sigma_Y} \tag{3.2}$$

Here, μ_X and μ_Y are the mean values for two matrices X and Y respectively. Cross-input neighborhood layer [5] computes the difference between feature maps produced by the convolution layers of two branches of the Siamese network. The output of normalized correlation and the cross-input neighborhood are fed to separate

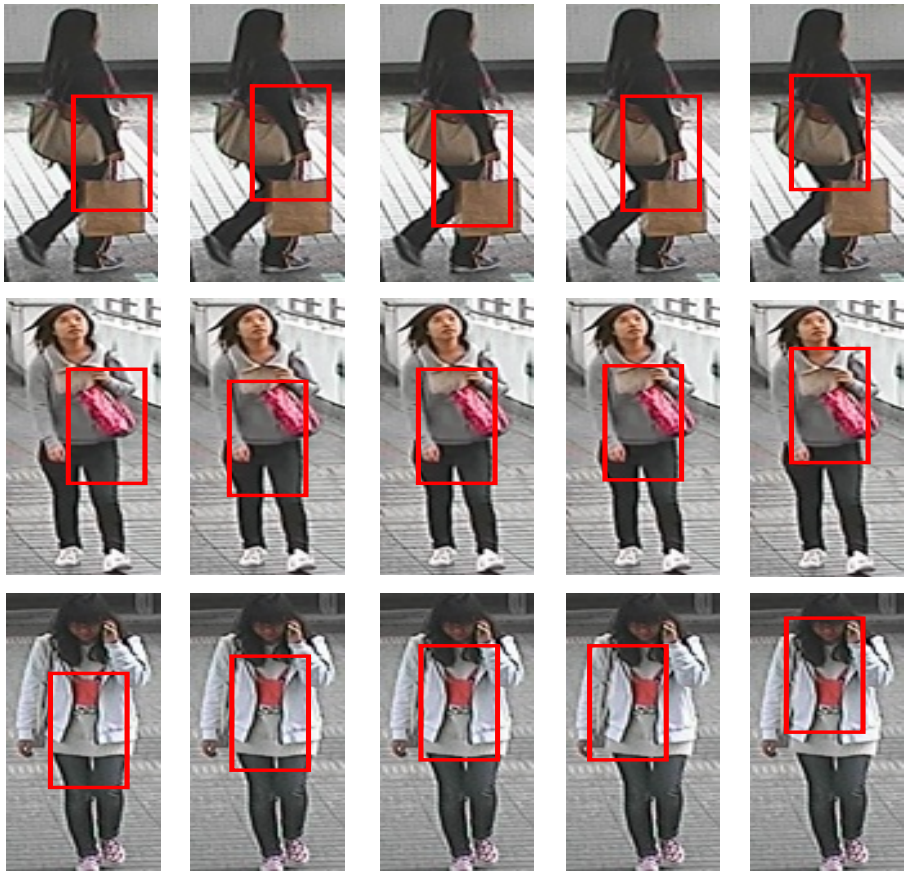

Figure 3.3: STN's localization behavior during training on CUHK01 dataset. Each row shows the localized image patch (in the red box) by STN for different training iterations. We find that STN converges to distinctive image regions after certain iterations.

cross patch feature aggregation layers which incorporate the contextual information and summarizes it. The feature aggregation layer is composed of two convolutions followed by max-pooling layer, and the output is 25 feature maps of size 5×17. The output feature maps are then fed to fully-connected layers of 500 hidden units. The fully-connected layers (one for patch image and another for global image) for each images are joined together with two softmax units. The output of the first softmax represents the likelihood that two images are same, and the other one represents the

likelihood that the images are different. To train our network, we use the standard cross-entropy loss and optimize the network parameters using the Stochastic Gradient Descent (SGD) algorithm.

Figure 3.2 shows the detailed architecture of our proposed network. Subramaniam et al. [7] also use FN for person re-identification. But the model in [7] computes the similarity of two input images only from the whole image features. In contrast, our proposed model first uses STN to localize the discriminative regions from the two input images, then the similarity score is computed based on these regions in addition to the whole images.

3.3 Experimental Evaluation

In this section, we firstly introduce the datasets used in our experiments (see Sec. 3.3.1). After that we describe network training strategies (see Sec. 3.3.2) and evaluation protocol (see Sec. 3.3.3). Finally we present our experimental results in Sec. 3.3.4.

3.3.1 Datasets

We conduct experiments on two benchmark datasets: CUHK01 [3] and CUHK03 [47] .

CUHK01 Dataset: This dataset consists of 3,884 images of 971 people [3]. For each person (or identity), there are 4 images captured from 2 different cameras. Following Subramaniam et al. [7], we conduct experiments in two different settings. In the first setting, we use 871 identities for training and the remaining 100 identities

for testing. In the second setting, we use 485 identities for training and the remaining 486 identities for testing.

CUHK03 Dataset: This is one of the largest benchmark dataset for person re-identification. It consists of 13,164 images of 1,360 pedestrians captured by 6 different surveillance cameras [47]. Each person is observed by 2 disjoint camera views. The dataset contains two different types of pedestrian bounding boxes – one as a result of manually labeling (referred as Labeled dataset) and the other that is algorithmically generated (referred as Detected dataset). In this work, we conduct experiments on both types. Again, we follow the experiment protocol of Subramaniam et al. [7] by randomly picking 1,260 identities for training and the rest for testing.

3.3.2 Network Training Strategies

We treat person re-identification as a binary classification problem. So we train the network using pairs of similar (i.e. positive pair) and dissimilar (i.e. negative pair) images. There exists data imbalance in the datasets – there are more negative pairs than positive pairs. Following previous work [7], we perform data augmentation to deal with the data imbalance. For every training set image of size $W \times H$, we sample several image patches (2 image patches for CUHK03 and 5 image patches for CUHK01 Dataset) around the image center and then apply random 2D translation drawn from a uniform distribution within the range of $[-0.05W, 0.05W] \times [-0.05H, 0.05H]$. This data augmentation strategy alleviates the training data imbalance issue across the datasets.

We implement our network using Torch 7 [50]. We train our network with mini-

batch of size 128. We use 0.9 as momentum and 0.05 as initial learning rate. Learning rate decay and weight decay are set to 1×10^{-4} and 5×10^{-4} respectively. We also fix the scaling value to 0.5 in the Spatial Transformer Network and learn translation parameters (t_x and t_y) only. Due to the data imbalance in most of the person re-identification dataset, after certain iteration the STN begins to consider whole image as patch. To mitigate this issue, we use fix scaling value to learn STN which gives better result along with the global image.

3.3.3 Evaluation Protocol

We present a comprehensive evaluation of our proposed method by comparing it with several state-of-the-art methods on CUHK01 and CUHK03 datasets. Following previous work, we rank the images present in the gallery image set based on the similarity with a probe image. Note that both the gallery images and the probe images are from test set. The intuition of this type of ranking is that the ground-truth matching gallary image should have the highest rank in the ideal case. In our experiments, we randomly select one image for each person/identity in the test set as a probe image and consider the remaining images as gallery images. For a probe image of a person, there is exactly one match in the gallery images. We perform 10 test trials on every probe image and report the averaged results in the tables along with several baselines. Note that the comparison with Subramaniam et al. [7], Ahmed et al. [5], and Li et al. [47] is of particular interest to us since they use similar deep learning architectures.

3.3.4 Results

CUHK01 Dataset: Table 3.1 and 3.2 summarize the experimental results on the CUHK01 dataset with 100 and 486 test identities. Our model outperforms the state-of-the-art method by nearly 5% in terms of the rank-1 accuracy. We believe that this performance gain is due to the discriminative regions learned by our network that is able to effectively distinguish between similar and dissimilar person images. Moreover, we train our network from scratch rather than pre-training it on a larger CUHK03 Labeled dataset, which is done by the state-of-the-art method in [7]. Note that the method in [7] is equivalent to our model without localizing the discriminative regions. Our model outperforms [7] by a large margin. This demonstrates the advantage of localizing discriminative regions in images for person re-identification.

Method	Rank-1	Rank-10	Rank-20
eSDC[43]	22.84	57.67	69.84
LDML[39]	26.45	72.04	84.69
KISSME[40]	29.40	72.43	86.07
Li et al.[47]	27.87	73.46	86.31
Ahmed et al.[5]	65.00	93.12	97.20
Wang et al.[51]	71.80	–	–
Subramaniam et al.[7]	81.23	97.39	98.60
Ours	**86.67**	**99.17**	**99.87**

Table 3.1: Performance of different methods at ranks 1, 10, and 20 on CUHK01 with 100 test IDs.

CUHK03 Dataset: Table 3.3 and 3.4 summarize the experimental results on the CUHK03 Labeled and Detected datasets, respectively. Our model outperforms the

Method	Rank-1	Rank-10	Rank-20
Mid-Level Filters [36]	34.30	65.00	74.90
Mirror-KFMA [52]	40.40	75.30	84.10
Ahmed et al.[5]	47.50	80.00	87.44
Ensembles [53]	51.90	83.00	89.40
CPDL[38]	59.50	89.70	93.10
Subramaniam et al. [7]	65.04	89.76	94.49
Ours	**71.35**	**93.08**	**96.80**

Table 3.2: Performance of different methods at ranks 1, 10, and 20 on CUHK01 with 486 test IDs.

state-of-the-art [7] method by nearly 2% in terms of the rank-1 accuracy. Figure 3.4 shows some qualitative retrieval results on this dataset.

Method	Rank-1	Rank-10	Rank-20
eSDC [43]	8.76	38.28	53.44
LDML [39]	13.51	52.13	70.81
KISSME [40]	14.17	52.57	70.03
Li et al. [47]	20.65	68.74	83.06
LOMO+XQDA [44]	52.20	92.14	96.25
Ahmed et al. [5]	54.74	93.88	98.10
LOMO+MLAPG [54]	57.96	94.74	98.00
Ensembles [53]	62.10	92.30	97.20
Subramaniam et al. [7]	72.43	95.51	98.40
Ours	**77.80**	**98.49**	**99.52**

Table 3.3: Performance of different methods at ranks 1, 10, and 20 on the CUHK03 Labeled dataset.

Figure 3.4: Qualitative retrieval results of our approach on CUHK03 dataset. The first column in each row represents a probe image. The remaining columns represent the retrieved results. The column highlighted in green is the ground-truth match.

Method	Rank-1	Rank-10	Rank-20
eSDC [43]	7.68	33.38	50.58
LDML [39]	10.92	47.01	65.00
KISSME [40]	11.70	48.08	64.86
Li et al. [47]	19.89	64.79	81.14
LOMO+XQDA [44]	46.25	88.55	94.25
Ahmed et al. [5]	44.96	83.47	93.15
LOMO+MLAPG [54]	51.15	92.05	96.90
Subramaniam et al. [7]	72.04	96.00	98.26
Ours	**74.48**	**96.16**	**98.28**

Table 3.4: Performance of different methods at ranks 1, 10, and 20 on the CUHK03 Detected dataset.

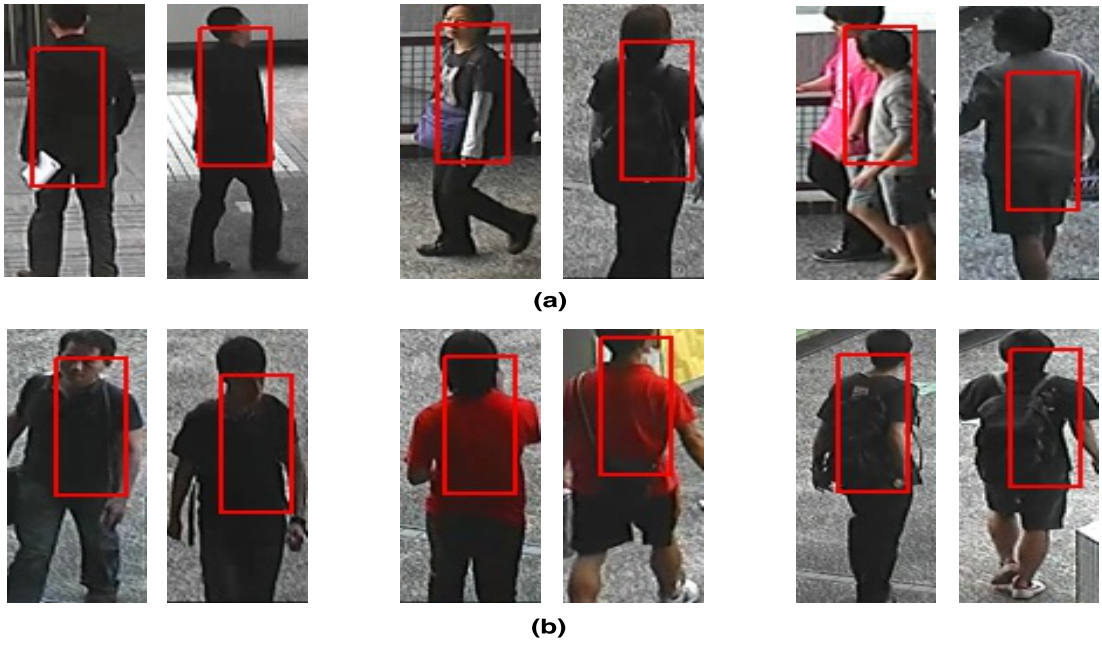

Figure 3.5: Some failure cases of our approach. (a) image pairs of the same person: our method incorrectly predict them as being dissimilar due to the lack of discriminative regions in these images; (b) image pairs of different persons: our method incorrectly recognize them as the same person, possibly because the localized discriminative regions in these image pairs have similar appearance.

Figure 3.5 shows some typical failure cases of our approach.

Chapter 4

Video-based Person Re-identification Using Refined Attention Networks

In this chapter, we propose a new attention based approach for re-identifying person from videos. Figure 4.1 shows the overall architecture of our proposed approach based on the Siamese network [34]. The input to the Siamese network is a pair of video sequences corresponding to the query video and the candidate video to be compared. The output of the Siamese network is a scalar value indicating how likely these two videos contain the same person. Each video goes into one of the two branches of the Siamese network. Each branch of the Siamese network is a Convolutional neural network used to extract the features of the input video. The parameters of two branches of the Siamese network are shared. Finally, the features from the two input videos are compared to produce the final output.

When a video goes through one of the two branches of the Siamese network, we first extract per-frame features on each frame of the input video. Then we compute an attention score on each frame indicating how important this frame is for the re-identification task. The intuition is that not all frames in a video are informative. The attention scores enable our model to ignore certain frames and only pay attention to informative frames in the video. The attention scores are then used to aggregate per-frame visual features weighted by the corresponding attention score to form a feature vector for the entire video sequence. We also propose an iterative refinement mechanism that uses the feature vector of the video to further refine the attention scores. Here the intuition is that the initial attention score of a frame is computed purely based on the frame. It does not take into account of other frames in the video. Since the feature vector of the entire video encodes contextual information of the whole video sequence, we can use this feature vector to further refine the attention scores. We can repeat this process for several iterations (see Sec. 4.5.4), where each iteration produces attention scores that focus more on the informative frames. Finally, the features of two input videos are compared to produce the output.

4.1 Frame-Level Features

Similar to [19], we extract frame-level features using both RGB color and optical flow channels. The colors contain information about the appearance of a person, while the optical flows contain information about the movement of the person. Intuitively, both of them are useful to differentiate the identity of the person. As a preprocessing step, we convert all the input images (i.e. video frames) from RGB to YUV color

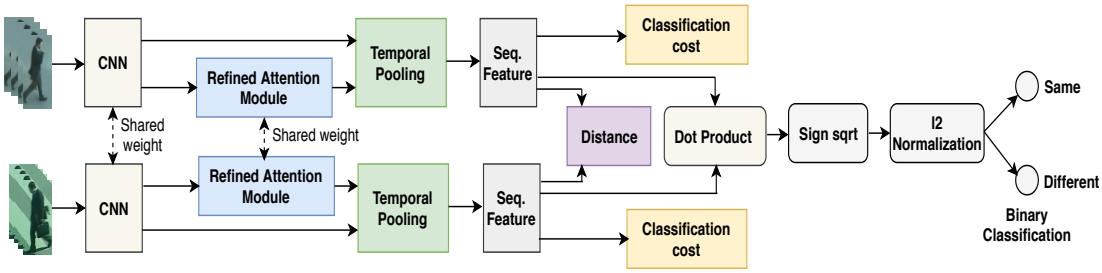

Figure 4.1: Overall architecture of our proposed Siamese network. It takes two input video sequences and pass to the Convolutional Neural Network (CNN) to extract features on each frame. The output from the CNN is fed to the attention module and generate an attention score for each frame. These attention scores combined with frame-level feature vectors to form a feature vector (i.e. temporal pooling) for the whole video. The video-level feature vectors are compared to decide whether the videos contain the same person

.

space. We normalize each color channel to have a zero mean and unit variance. The Lucas-Kanade algorithm [55] is used to calculate both vertical and horizontal optical flow channels on each frame. We resize each frame to have a spatial dimension of 56×40. The optical flow field F of the frame is split into two scalar fields F_x and F_y corresponding to the x and y components of the optical flow. In the end, each frame is represented as a $56 \times 40 \times 5$ input, where the 5 channels correspond to 3 color channels (RGB) and 2 optical flow channels (x and y).

We fine-tuned CNN architecture of [19] to extract frame-level features for an input video. Note that we replace the fully connected in the end by two new fully connected layers that produce 1024 and 128 dimensional feature vectors respectively. Given an input video with T frames, we apply the CNN model in Fig. 4.2 on each frame of the input video. In the end, each frame $\mathbf{x}_i$ $(i = 1, 2, ..., T)$ is represented as a 128 dimensional feature vector, i.e. $\mathbf{x}_i \in \mathbb{R}^{128}$.

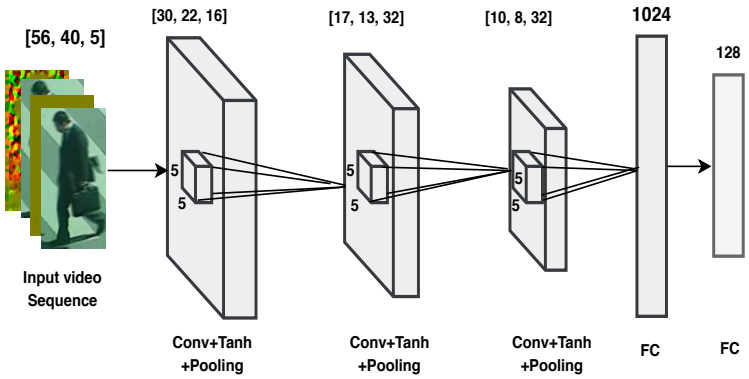

Figure 4.2: Our ConvNet architecture for extracting frame-level features. The ConvNet process a frame (both RGB channels and optical flow channels) using a series of layers. Each layer is composed of convolution, maxpooling and hyperbolic-tangent (Tanh) activation-function. The convolution uses 5x5 kernel with 1x1 stride and 4x4 zero padding. The output from the third convolution layer is fed to two fully connected layers which generate feature vectors of length 1024 and 128 respectively to represent this frame.

4.2 Temporal Attention Network

Motivated by the recent success of attention based models [25; 56; 27; 57], we propose an attention based approach for re-identifying person from videos. The intuition behind the attention based approach is inspired by the human visual processing [21]. Human brains often pay attention to different regions of different sequences when trying to re-identify persons from videos. Based on this intuition, we propose a deep Siamese architecture where each branch generates attention scores of different frames based on the frame-level CNN features. The attention score of a frame indicates the importance of this frame for the re-identification task.

As shown in Figure 4.1, each input video sequence (sequence of frames with optical flow) is passed to the CNN to extract frame-level feature maps. Using fully connected layers, CNN generates feature vector for each video frame. The sequence of feature

vectors are passed to the attention network to generate attention scores. More specifically, for each feature vector $\mathbf{x}_i$ corresponding to the i-th frame, we compute an attention score α_i indicating the importance of this frame. The attention score is obtained by applying a linear mapping followed by a sigmoid function. Here,we are use the same parameters for the linear mapping on all frames. Let θ be the vector of parameters for the linear mapping. Now the attention score α_i is calculated using the following equations:

$$z_i = \theta^T \mathbf{x}_i \tag{4.1a}$$

$$\alpha_i = \frac{1}{1 + \exp(-z_i)}, \quad \text{where } i = 1, 2, ..., T \tag{4.1b}$$

We have also tried using softmax instead of sigmoid function in Eq. 4.1 and found that it does not perform as good as the sigmoid function. Previous work [58] has made similar observations. Once we have obtained an attention score α_i for each frame in the video, we can then combine the attention scores α_i $(i = 1, 2, ..., T)$ with frame-level feature vectors to create a weighted feature vector $\mathbf{f}$ as follows:

$$\mathbf{f} = \sum_{i=1}^{T} \alpha_i \mathbf{x}_i, \quad \text{where } i = 1, 2, ..., T \tag{4.2}$$

where $\mathbf{f}$ can be seen as a feature vector for the entire video which takes into account the importance of each frame in the video.

4.3 Attention Refinement

In principle, we can directly use the video-level feature vector in Eq. 4.2 for person re-identification, e.g. by comparing the feature vectors of two videos. But one possible

limitation is that the attention score in Eq. 4.1 is calculated on each frame in the video separately. In other words, the attention scores for frames in a video are independent of each other. This is not very intuitive – the attention score of a frame should depend on the visual information of the video, which in turn depends on all frames in the video. In this section, we introduce a strategy to refine the attention scores so that they are all coupled together in the end. In the experiment section, we will show that this attention refinement improves the performance of our model.

The basic idea of the attention refinement is to use the video-level feature vector $\mathbf{f}$ (Eq. 4.2) as one of the input to re-compute the attention score on each frame in the video. Since the video-level feature vector $\mathbf{f}$ depends on all frames in the video, the new attention score on a frame will implicitly depend on all frames in the video as well. The new attention scores can then be used to update the video-level feature vector. This process can be repeated for multiple iterations. Let us define α'_i as to be the new attention score. In this work, we simply concatenate $\mathbf{f}$ to each frame-level feature $\mathbf{x}_i$, then apply a linear mapping as follows:

$$z'_i = \theta'^T \text{concat}(\mathbf{x}_i, \mathbf{f}) \tag{4.3a}$$

$$\alpha'_i = \frac{1}{1 + \exp(-z'_i)}, \quad \text{where } i = 1, 2, ..., T \tag{4.3b}$$

where concat$(\cdot)$ means the concatenation of two vectors. Then the new video-level feature vector $\mathbf{f}'$ can be computed as:

$$\mathbf{f}' = \sum_{i=1}^{T} \alpha'_i \mathbf{x}_i, \quad \text{where } i = 1, 2, ..., T \tag{4.4}$$

We alternate between updating attention scores (Eq. 4.3) and updating video-level feature vector (Eq. 4.4) for several iterations. Empirically, we have found 3 iterations

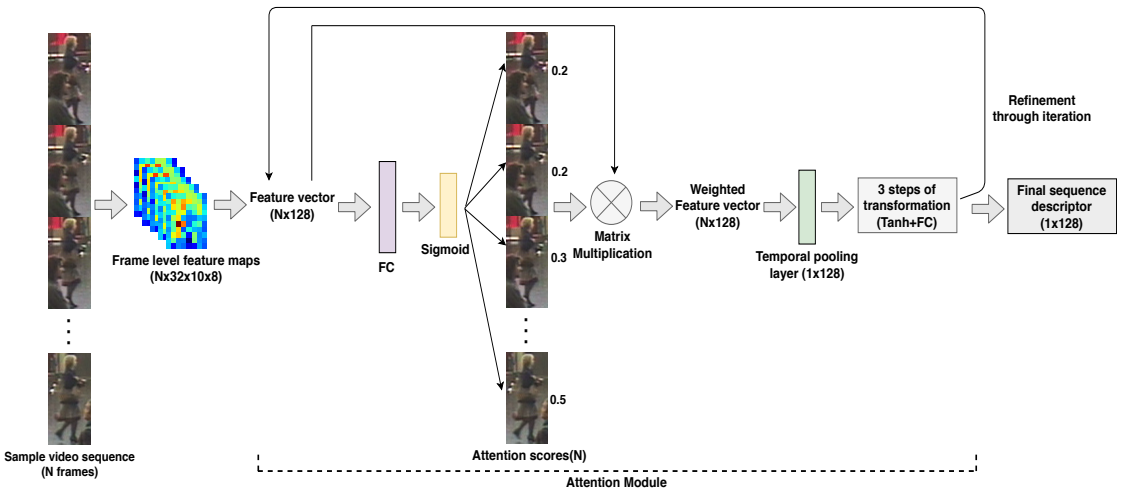

Figure 4.3: Illustration of our proposed refined attention network architecture. The input is a feature matrix of dimensions $N \times d$ where N is the number of frames in the sequence and d is the dimension of frame-level features. We generate N attention scores by applying linear mapping on the feature vectors followed by a sigmoid function. These attention scores are combined with frame-level features via temporal pooling to form a feature vector for the entire video. We use the video-level feature vector as one of the inputs to further refine the attention score on each frame. We then compute a new video-level feature vector using the new attention scores.

give the best performance (see Sec. 4.5.4). Figure 4.3 shows the architecture of this attention refinement.

4.4 Model Learning

Our model is a form of the Siamese network (Fig. 4.1). It has two identical branches with shared parameters. The detail architecture of each branch is shown in Fig. 4.3. Each branch takes a video as its input and produces a feature vector of the video according to Eq. 4.4. Let $\mathbf{f}'_1$ and $\mathbf{f}'_2$ be the feature vectors of the two input videos to the Siamese network. We use Y_1 and Y_2 to denote the identity of

the person in these two videos. Similar to [19; 21], we calculate Euclidean distance between these two feature vectors and use the following squared hinge loss(H_{loss}) as the loss function to train our network:

$$\mathcal{L}_{hinge} = \begin{cases} \frac{1}{2}\|\mathbf{f}_1' - \mathbf{f}_2'\|^2, & Y_1 = Y_2 \\ \frac{1}{2}[\max(0, m - \|\mathbf{f}_1' - \mathbf{f}_2'\|)]^2, & Y_1 \neq Y_2 \end{cases} \tag{4.5}$$

where m is a hyper-parameter that represents the margin of separating the two classes in $\mathcal{L}_{hinge}$. By minimizing this squared hinge loss, the distance between feature vectors will be small if the two videos contain the same person (i.e. $Y_1 = Y_2$). The distance will be large if the two videos contain two different persons (i.e. $Y_1 \neq Y_2$).

We also use a standard binary cross-entropy ($\mathcal{L}_{sim}$) that classifies the input videos to be same or different. For this, we firstly compute the inner product I of the video features and then perform a signed square-root step (i.e. $s \leftarrow sign(I)\sqrt{|I|}$). The resulting output is followed by a l_2 normalization ($N \leftarrow \frac{s}{\|s\|_2}$) and a softmax operation.

Following [19], we add an additional loss in each of the two branches of the Siamese network to predict the person's identity. Each branch uses the feature vector for the input video extracted from the network and applies a linear classifier to predict one of the K identities of the person. We use the softmax loss for the person identification classification. Let $\mathcal{L}_{id1}$ and $\mathcal{L}_{id2}$ be the loss functions of the two branches. The final loss function is the combination of the two identify classification losses, similarity loss and the squared hinge loss.

$$\mathcal{L}_{final} = \mathcal{L}_{id1} + \mathcal{L}_{hinge} + \mathcal{L}_{sim} + \mathcal{L}_{id2} \tag{4.6}$$

The network is trained end-to-end by optimizing the loss function in Eq. 4.6 using stochastic gradient descent. Following [19], we remove both classification losses, the squared hinge loss and similarity loss from the network after training is done. During testing, we only use the feature vectors generated by the two branches of the Siamese network and directly compare their distance for re-identification.

4.5 Experimental Evaluation

In this section, we firstly introduce the datasets used in our experiments (Sec. 4.5.1). We then describe the experimental setup and some implementation details (Sec. 4.5.2). We present the results of experiment in Sec 4.5.3 and Sec 4.5.4.

4.5.1 Datasets

We conduct experiments on three benchmark datasets: iLIDS-VID [20], PRID-2011 [59] and MARS [60].

iLIDS-VID Dataset: This dataset consists of video sequences of 300 persons where each person is captured by a pair of non-overlapping cameras. The length of each video sequence varies from 23 to 192 frames with an average of 73 frames. The dataset is quite challenging due to lot of occlusions, illumination changes, background clutters and so on.

PRID-2011 Dataset: This dataset contains video sequences of 749 persons. For the first 200 persons (or identities), there are two video sequences captured by two different cameras. The remaining persons appear in only one camera. Each sequence contains between 5 to 675 frames, with an average of 100 frames. Compared

with iLIDS-VID, the PRID-2011 dataset contains less occlusions since the videos are captured in a relative simple environment.

MARS Dataset: The Motion Analysis and Re-identification Set (MARS) is the largest video-based person re-identification dataset that contains 1,261 different pedestrians. Each pedestrian is captured by at least two cameras. DPM detector and GMMCP tracker are used to generate the tracklets. There are, on average, 13.2 tracklets for each pedestrian. Table 4.1 shows the summary of these three benchmark datasets.

Dataset	iLIDS-VID	PRID-2011	MARS
Total no. of id.	300	749	1,261
No. id in multiple cameras	300	200	1,261
No. track-lets	600	400	21K
No. of boxes	44K	40K	1M
Image resolution	64x128	64x128	128x256
No. of camera	2	2	6
Detection procedure	hand	hand	algorithm
Evaluation metric	CMC	CMC	CMC

Table 4.1: Summary of basic information of the three datasets used in our experiments.

4.5.2 Setup and Implementation Details

We follow the experiment protocol of McLaughlin et al. [19]. On each of the two datasets (iLIDS-VID and PRID-2011), we randomly split the dataset into two equal subsets where one subset is used for training and remaining one for testing. For

evaluating our proposed method, we use the Cumulative Matching Characteristics (CMC) curve which is a ranking based evaluation metric. In the ideal case, the ground-truth video sequence should have the highest rank. For each dataset, we repeat the experiment 10 times and report the average result over these 10 runs. In each run, we randomly split the dataset into training/test sets. Standard data augmentation techniques, such as cropping and mirroring, are applied to increase the amount of training data. We initialize the weights in the network using the initialization technique in [61]. For training our network, we consider equal numbers of positive and negative samples. We set the margin in the hinge loss (Eq. 4.5) as $m = 2$. The network is trained for 1000 epochs with a batch size of one. The learning rate in the stochastic gradient descent is initially set to be $1e^{-3}$. We decrease the learning rate by a factor of 10 after 300 and 600 on the PRID-2011 dataset. Due to the variable-length of video sequences in both datasets, we use sub-sequences of 16 consecutive frames ($T = 16$) during training. Sometimes, this length is greater than the real sequence length. In that case, we consider the whole set of images (frames) as the sub-sequence. A full epoch consists of a pair of positive and negative sample. During testing, we consider a video sequence captured by the first camera as the probe sequence and a video sequence captured by the second camera as a gallery sequence. We use at most 128 frames in a testing video sequence. Again, if the length is greater than the real sequence, we consider the whole set of images as the video sequence. Similar strategies have been used in previous work [19]. For the MARS dataset, we follow the experimental protocol of state-of-the-art method by Xu et al. [21].

4.5.3 Results

We present the results on the three benchmark datasets and compare with other state-of-the-art methods in Table 4.2 and Table 4.4. From the CMC rank, we see that our method with attention refinement outperforms all other state-of-the-art methods by nearly 2% and 3% in terms of rank-1 accuracy on the iLIDS-VID and PRID-2011 dataset, respectively. On the MARS dataset, we outperform the state of the art by a big margin of 18% on rank-1 accuracy. Figure 4.4 shows some qualitative retrieval results after applying our proposed method on the challenging iLIDS-VID dataset.

iLIDS-VID				
Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**64**	**88**	**96**	**98**
Xu et al. [21]	62	86	94	98
Zhou et al. [23]	55.2	86.5	-	97.0
McLaughlin et al. [19]	58	84	91	96
Yan et al. [22]	49.3	76.8	85.3	90.1
STA [18]	44.3	71.7	83.7	91.7
VR[20]	35	57	68	78
SRID[62]	25	45	56	66
AFDA[63]	38	63	73	82
DTDL[64]	26	48	57	69

Table 4.2: Comparison of our proposed approach with other state-of-the-art methods on the iLIDS-VID dataset in terms of CMC(%) at different ranks.

PRID-2011				
Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**82**	**97**	**99**	99
Xu et al.[21]	77	95	99	99
Zhou et al.[23]	79.4	94.4	-	**99.3**
McLaughlin et al.[19]	70	90	95	97
Yan et al. [22]	58.2	85.8	93.7	98.4
STA [18]	64.1	87.3	89.9	92
VR[20]	42	65	78	89
SRID[62]	35	59	70	80
AFDA[63]	43	73	85	92
DTDL[64]	41	70	78	86

Table 4.3: Comparison of our proposed approach with other state-of-the-art methods on the PRID-2011 dataset in terms of CMC(%) at different ranks.

Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**62**	**85**	**93**	**95**
Xu et al.[21]	44	70	74	81
McLaughlin et al. [19] (obtained from [21])	40	64	70	77

Table 4.4: Comparison (CMC(%)) of our proposed approach with previous methods on the MARS dataset.

4.5.4 Effect of Iterative Refinement

We conduct empirical study on the training set of the iLIDS-VID and MARS dataset to analyze the effect of the attention refinement (i.e. number of iterations) on the overall performance of the proposed network. We randomly divide the training

Figure 4.4: Qualitative retrieval results of our proposed method on the challenging iLIDS-VID dataset. The first column represents the probe video sequence. The remaining columns correspond to retrieved video sequences sorted by their distances to the probe video sequence. Here, we use a single image to represent each retrieved video sequence. The green boxes indicate the ground-truth matches. We can see that the ground-truth matches are ranked very high in the list.

data of iLIDS-VID as well as MARS into two parts: one for learning the model parameters and the other one for validation. We select 110 persons for training the model and the remaining 40 persons for validation from iLIDS-VID dataset. For MARS dataset, we select 400 identities for training and the remaining 225 identities for validation purpose. We train the model on the training videos and report the performance (CMC(%)) on the validation set for different number of iterations in Table 4.5 and Table 4.6. We observe that the performance gradually improves until iteration 3. After that, the performance starts to drop. Based on this empirical result,

we choose 3 iterations in our experiments.

iLIDS-VID				
# iterations	Rank-1	Rank-5	Rank-10	Rank-20
0 (No iteration)	60	92	97	100
1 (1 iteration)	70	95	97	100
2 (2 iterations)	62	95	97	100
3 (3 iterations)	**77**	**97**	**97**	**100**
4 (4 iterations)	70	97	97	100
5 (5 iterations)	65	97	97	100

Table 4.5: Validation performance for different number of iterations on the iLIDS-VID dataset.

MARS				
# iterations	Rank-1	Rank-5	Rank-10	Rank-20
0 (No iteration)	56	80	87	89
1 (1 iteration)	57	80	87	89
2 (2 iterations)	56	80	86	90
3 (3 iterations)	56	**80**	**88**	**90**
4 (4 iterations)	**58**	78	87	90
5 (5 iterations)	58	79	87	89

Table 4.6: Validation performance for different number of iterations on the MARS dataset.

Chapter 5

Conclusion and Future Work

In this thesis, we have presented deep learning based methods to re-identify person from images and videos. We have proposed two novel approaches in the field of person re-identification. First, we proposed an end-to-end deep neural network architecture that localizes discriminative image regions for person re-identification. The novelty of this thesis is that firstly it localizes discriminative salient image regions and then computes the similarity based on these image regions in conjunction with the whole image. Second, we have proposed an attention-based deep architecture for video-based re-identification. The attention module calculates frame-level attention scores, where the attention score indicates the importances of a particular frame. The output of the attention module can be used to produce a video-level feature vector which can be refined iteratively to generate rich feature information. In addition, we show that this attention refinement increase the performance of our proposed method.

Currently, our image based person re-identification network only localizes one discriminative region in each image. We plan to extend our model to localize multiple

discriminative regions by using more than one Spatial Transformer Network in the model which can be an interesting and important direction for future work. The intuition is that a person can posses multiple discriminative regions which can be useful for re-identification. For each discriminative region, we have to add a Siamese network where each branch will share parameters and extract discriminative region. The pair of discriminative region will feed to the fused network and finally generate two similarity scores representing a measurement of how similar two images are. Moreover, in future we can also consider spatial attention with the conjunction of temporal attention for video based re-identification.

Bibliography

[1] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996. 1

[2] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 356–370, 2017. 2

[3] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 31–44. 2, 8, 18

[4] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672. 2

[5] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916. 2, 8, 9, 16, 20, 21, 22, 23

[6] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016. 2

[7] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Advances in Neural Information Processing Systems*, 2016, pp. 2667–2675. 2, 8, 9, 15, 18, 19, 20, 21, 22, 23

[8] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025. 2, 14, 15

[9] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *IEEE International Conference on Computer Vision*, 2015. 3, 8

[10] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *IEEE International Conference on Computer Vision*, 2017. 3, 8

[11] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," in *IEEE International Conference on Advanced Video and Signal based Surveillance*, 2017. 3, 8

[12] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*, 2016. 3, 8

[13] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 8

[14] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for person re-identification," in *IAPR International Conference on Pattern Recognition*, 2014. 3, 8

[15] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 8

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Salable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015. 3, 8

[17] Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, and Q. Tian, "Video-based person re-identification by deep feature guided pooling," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 3, 10

[18] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *IEEE International Conference on Computer Vision*, 2015. 3, 10, 36, 37

[19] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional neural network for video-based person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 10, 26, 27, 32, 33, 34, 35, 36, 37

[20] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014. 3, 10, 33, 36, 37

[21] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *IEEE International Conference on Computer Vision*, 2017. 3, 4, 10, 28, 32, 35, 36, 37

[22] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*, 2016. 3, 10, 36, 37

[23] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 4, 5, 10, 11, 36, 37

[24] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *International Joint Conference on Artificial Intelligence*, 2016. 3, 10

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015. 3, 10, 28

[26] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 10

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015. 3, 10, 28

[28] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008. 7

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal oc-

curence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 7

[30] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision Workshop*, 2012. 7

[31] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[32] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 7

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012. 7

[34] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 8, 25

[35] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367. 8

[36] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151. 8, 22

[37] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency weighted features for person re-identification," in *European Conference on Computer Vision*. Springer, 2014, pp. 191–208. 8

[38] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification." in *International Joint Conference on Artificial Intelligence*, 2015, pp. 2155–2161. 8, 22

[39] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *IEEE International Conference on Computer Vision*, 2009, pp. 498–505. 8, 21, 22, 23

[40] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295. 8, 21, 22, 23

[41] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *IEEE International Conference on Image Processing*, 2013, pp. 3567–3571. 8

[42] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573. 8

[43] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593. 8, 21, 22, 23

[44] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206. 8, 22, 23

[45] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5645–5658, 2015. 8

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *International Conference on Pattern Recognition*, 2014, pp. 34–39. 8

[47] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159. 8, 9, 18, 19, 20, 21, 22, 23

[48] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 753–769. 10

[49] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 539–546. 12

[50] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, Neural Information Processing Systems Workshop*, 2011. 19

[51] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296. 21

[52] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification." in *International Joint Conference on Artificial Intelligence*. Citeseer, 2015, pp. 3402–3408. 22

[53] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855. 22

[54] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693. 22, 23

[55] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *International joint conference on Artificial intelligence*, 1981. 27

[56] C. N. dos Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *Computing Research Repository*, 2016. 28

[57] W. Yin, H. Schutze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural networks for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, 2016. 28

[58] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 29

[59] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conference on Image Analysis*, 2011. 33

[60] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016. 33

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into recitifers: Surpassing human-level performance on imagenet classification," in *IEEE International Confernece on Computer Vision*, 2015. 35

[62] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *IEEE Conference on Computer Vision and Patten Recognition Workshop*, 2015. 36, 37

[63] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *British Machine Vision Conference*, 2015. 36, 37

[64] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *IEEE International Conference on Computer Vision*, 2015. 36, 37