Modality Guided Representation Learning for Person Re-identification

By

Kajal Kansal

under the supervision of

Dr. A.V. Subramanyam

Indraprastha Institute of Information Technology-Delhi

May, 2020

©Indraprastha Institute of Information Technology-Delhi, New Delhi, 2020

Modality Guided Representation Learning for Person Re-identification

By

Kajal Kansal

submitted

in partial fulfillment of the requirements

for the award of the degree of

Doctor of Philosophy

to the

Indraprastha Institute of Information Technology-Delhi

Okhla Industrial Estate, Phase III

New Delhi, India - 110020

May, 2020

Indraprastha Institute of Information Technology-Delhi
Okhla Industrial Estate, Phase III
New Delhi-110020, India
May-2020

Certificate

This is to certify that the thesis titled **Modality Guided Representation Learning for Person Re-identification** being submitted by **Kajal Kansal** to the Indraprastha Institute of Information Technology-Delhi, for the award of the degree of **Doctor of Philosophy**, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. A.V. Subramanyam
Associate Professor
Indraprastha Institute of Information Technology
New Delhi-110020
May, 2020

Indraprastha Institute of Information Technology-Delhi
Okhla Industrial Estate, Phase III
New Delhi-110020, India
May-2020

Declaration

This is to be certified that the dissertation entitled **Modality Guided Representation Learning for Person Re-identification** being submitted by **Kajal Kansal** to the **Indraprastha Institute of Information Technology-Delhi**, for the award of degree of **Doctor of Philosophy**, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Dr. A.V. Subramanyam**.

The study pertains to this dissertation has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

Kajal Kansal
PhD student
IIIT-Delhi, India

Acknowledgement

I would like to express my appreciation to my supervisor **Dr. A.V. Subramanyam** for giving me the opportunity to work on this thesis and enlighten my PhD journey through his continuous support, active monitoring, and timely feedback. I wish to thank him for his co-operation, enthusiasm, highest priority to research, and patience throughout my Ph.D. Without his valuable thoughts and recommendations, It would have been difficult for me to complete this dissertation. I would like to thank him for providing excellent research facilities.

I am grateful to the **Indraprastha Institute of Information Technology-Delhi** for providing excellent infrastructure, research environment and research oriented course work. I would like to express gratitude to **Visvesvaraya PhD Scheme** for providing me with a research fellowship that helped me financially throughout my PhD.

My sincere thanks also goes to Professor Shin'ichi Satoh for offering me the internship opportunities in his group at NII, Tokyo, Japan. This one year internship has been a special part of my PhD time and nourished my research skills and brought a big positive change in me. I would like to say a big thanks to Dr. Zheng Wang for his guidance, motivation and positive words throughout this internship. I thank him for the continuous technical/nontechnical discussions. He has been a greatest role model for my research during the internship.

I would like to thank all my spiritual teachers namely Dr. Satish Gupta, Dr. Mizue Honda, Dr. Katheline, Dr. Sachin Gupta, Dr. Mohit Gupta and Dr. Ramesh Sundram to motivate me through their angelic lives and make me a divine being with good management skills, positivity, emotionally independent and energetic through their spiritual guidance and meditation sessions. I would also like to thank my yoga teacher Ajay Saxena for his yoga classes and encourage me to keep myself physically fit and mentally confident.

I would like to thank all my fellow colleagues namely Kunal Saini, Ramesh Srivastava, Shishir Sharma, Nishima Arora and Rahul Duggal for the discussions of deep learning related stuff. I would like to thank Rishi Gargi and Yang Fan to discuss several computer and server related issues. I would like to thank Shivangi Singhal, Monika Jain and Astha Verma for discussion of several course and research work. I also wish to thank Dr. Hemant K. Aggarwal, Siddarth Dawar, Ramneek Kaur and Ambuj Mehrish for their feedbacks in the presentations. I also wish to thank Dr. Naushad Ansari to encourage me through several discussions and motivates me through his achievements.

I would also like to gratefully acknowledge to Nidhi Goyal, Neelam Sharma, Anjali Lathwal, Tania Sidana, Sarita Poonia, Anjali Dhall, Sumit Patiyal, Neetesh Pandey and Anand Singh for their invaluable friendship and emotional support.

I am extremely grateful to my committee members and collaborators : Dr. Rajiv Ratan Shah, Dr. Tanmoy Chakraborthy, Prof. Mohan Kankanhalli and Dr. Dilip K. Prasad for their valuable comments. My gratitude also extends to the technical and admin support staff of IIIT-D, especially Mr Adarsh and Mrs Priti for being extremely helpful in the fast resolution of all technical

and admin related matters.

My family's support, love and co-operation has been enormous throughout my journey. I am thankful to my parents for their blessings, brother, sister and all my relatives for their full support under all conditions not just during Ph.D. but throughout my life.

To my dear Supreme Shivbaba, Although there are no worldly words to express the gratitude to the most special being of my life. I would like to dedicate my pure feelings to express thanks for giving me the companionship, powers, blessings, recharging, enlightenment and pure energy throughout my life. There is a special place in my soul for you. I dedicate this thesis to you.

Kajal Kansal

Contents

List of Figures

Page

List of Tables

Page

Abstract

The rise of surveillance cameras has led to a significant focus on large scale deployment of intelligent surveillance systems. Person re-identification (Re-ID) is one of the quintessential surveillance problems. Person Re-ID is the task of matching people from the non-overlapping multi-camera network. It is a non-trivial problem because of the presence of several visual recognition challenges such as pose change, occlusion, illumination variation, low resolution, and additional challenges due to temporal dimension in videos for homogeneous Re-ID. In case of heterogeneous Re-ID, the large modality difference is a big challenge. There are various applications such as long-term multi-camera tracking, crime prevention, forensic search, threat detection, instance search, activity analysis, photo-tagging, and many more intelligent applications.

In this dissertation, we propose homogeneous and heterogeneous Re-ID models. In homogeneous case, we contribute towards video-video and image-to-image matching. In heterogeneous Re-ID, we propose novel models for RGB-IR and text-image matching.

Our first work is based on video-video Re-ID. In this work, we propose a novel shallow end-to-end model. It incorporates two stream CNNs, discriminative visual attention and recurrent neural network with triplet and softmax loss to learn the spatio-temporal fusion features and improve the generalization ability. In addition, we contribute a large novel dataset of air borne videos for person Re-ID, named DJI01. It includes various challenging conditions like occlusion, illumination-changes, people with similar clothes and same people at different days. The elaborate qualitative and quantitative experiments on PRID-2011 [1], iLIDS-VID [3], MARS [2] and our drone dataset DJI01 demonstrate the robustness of the extracted discriminative features and efficacy of the proposed model.

The second work aims at developing the image-based Re-ID model. In order to obtain strong generalizable as well as discriminative features, we propose a novel deep reconstruction re-identification network (HDRNet). HDRNet comprises of an encoder and a multi-resolution decoder, which can learn embeddings invariant to pose, occlusion, illumination, and low-resolution. We further propose a hybrid sampling strategy to boost the effectiveness of the training loss function. In addition, we propose test set augmentation using reconstructed images to explicitly transform single query to multi query setting. In our multi-tasking approach, the feature robustness is enhanced by the multi-resolution decoder and the overall performance is further improved by sampling strategy and test data augmentation. The rigorous analysis on publicly available datasets CUHK03 [7], Market-1501 [8] and DukeMTMC-reID [9] demonstrate the state-of-the-art accuracy.

Homogeneous Re-ID assumes single modality which restricts its utility under scenario where samples are captured under different spectrum. We address the limitations of homogeneous Re-ID models in our second contribution and propose an image-based heterogeneous Re-ID model.

Visible-infrared (RGB-IR) Re-ID is one of the important heterogeneous Re-ID tasks for surveillance applications under poor illumination conditions. In addition to conventional Re-ID challenges, the spectrum discrepancy scales up the difficulty of the problem. In order to address this, we propose to disentangle the spectrum information while learning the identity

discriminative features. To extract these features, we propose a novel network with disentanglement loss which can distill identity features and dispel spectrum features. Our network has two branches, spectrum dispelling and spectrum distilling branch. On spectrum dispelling branch, we apply identification loss to learn the identity related and spectrum disentangled features. On spectrum distilling branch, we apply an identity-dispeller loss to fool the identity classifier so that it primarily learns spectrum related information. The entire network is trained in an end-to-end manner, which minimizes spectrum information and maximizes invariant identity relevant information at spectrum dispelling branch. Extensive experiments on existing datasets SYSU-MM01 [5] and RegDB [6] demonstrate the superior performance of our approach.

The previous problem settings deal with images or videos only. However, in many cases, there might be only a verbal description of the persons' appearance available. These descriptions can also be used as the cues for finding a specific person from visual surveillance data. Such a scenario motivates us to address new challenges beyond the single and multi-modal visual Re-ID, named as text-image Re-ID. To this end, we first analyze the major challenges of text-image Re-ID which include *text complexity* arising due to different words with the same meanings and *alignment uncertainty* occurring during matching due to poor correspondence of text-image pairs. To solve these challenges, we propose an end-to-end Hierarchical Attention Alignment Network. Our model comprises of: i) a new strategy of Term Frequency-Inverse document Frequency thresholding to extract the salient tokens to alleviate the challenge of *text complexity*; ii) a hierarchical attention alignment network to determine the potential relationships of image content and textual information at different levels, namely, word-patch level, phrase-patch level, and sentence-image level for addressing *alignment uncertainty*. Since hierarchical attention exploits salient regions has an additional advantage for performing retrieval in fine-grained text-image matching. The network is optimized via joint weighted hierarchical attention loss and cross-modal loss in an end-to-end manner. Extensive quantitative and qualitative analysis on the challenging datasets CUHK-PEDES [10] and Flickr-30K [11] demonstrate the superiority of the proposed approach.

Chapter 1

Introduction

1.1 Surveillance

Surveillance and security cameras are continuously growing day by day in our society. The research community has actively focused on investigating automatic methods of surveillance for several decades. In video surveillance, one of the key tasks is to detect, identify, and monitor person in crowded and public scenes such as airports, train stations, and supermarkets. In the video surveillance context, detecting the whereabouts of humans is the first requirement if any event involving people has to be detected. The problem of locating a person in the surveillance footages from overhead view has been actively researched since last few decades. Automated person detection finds its applications in many areas including human robot interaction, surveillance, pedestrian protection systems, automated image and video indexing, as it provides the fundamental information for semantic understanding of the CCTV video footages. Another important task is Person Identification which identifies an individual uniquely. An individual can be uniquely identified by using unique personal identification number (PIN) and also using biometrics like fingerprint, face etc. The development of accurate and efficient person identification methods is a major area of research in the computer vision, biometric, surveillance, and security communities. Amongst major surveillance tasks, person re-identification (Re-ID) is one of the critical problems. Person Re-ID [12–18] is the field of study that deals with matching of people across disjoint camera views. More specifically, re-identification of an individual is the task of matching a person in diverse scenes, obtained from different cameras distributed over non-overlapping views. It has gained much attention in the recent past due to its importance in tasks such as long-term multi-camera tracking, threat detection, instance search, forensic search, crime prevention, activity analysis, photo-tagging, trajectory generation of person-of-interest, anti-terrorism and many more intelligent applications in real world. The main focus of Re-ID is to develop an algorithm such that if a person is moving from one camera to another camera then the algorithm should be able to compare the query (image/video/text) against all the gallery (image/video) and identify the correct match. However, identifying a person from the appearance often becomes a difficult task under uncontrolled camera and environmental variations. Thus describing a person in terms of robust representation is a challenging task. A typical person re-identification scenario can be seen in Figure 1.1.

More formally, Re-ID can be defined as follows. Let the training data be $X \in R^{d \times N}$. Each column of the data matrix X, $x_i \in R^{d \times 1}$, represents the i^{th} training sample. Let $y_i \in R^{K \times 1}$ be the one-hot encoded label of x_i where K is the number of identities. Further, let D be a distance measure such as Euclidean distance. Then, the distance between a pair of different samples, x_i

Figure 1.1: *A Typical Camera Network System*

and x_j, can be computed as,

$$D(x_i, x_j) = \|x_i - x_j\|_2^2 \tag{1.1}$$

We aim to learn an optimal deep feature representation model for Re-ID such that the sample x_j whose distance is minimum to the sample x_i of the same identity is the closest match and $D(x_i, x_j) < D(x_i, x_j')$, where x_j' is a sample of any other person.

1.2 Homogeneous and Heterogeneous Re-ID

Person Re-ID algorithms can be broadly classified under two categories "Homogeneous Re-ID" and "Heterogeneous Re-ID". Homogeneous Re-ID algorithms for video-video [19–21] and image-image [2, 8, 22] have been proposed for matching the person under the same modality. A typical model of the homogeneous Re-ID [12, 23–27] system takes a same modal pair of probe and gallery data as an input to the network and outputs a similarity score between the two which depicts whether the pair belongs to the same person or not. Colour information is one of the most important appearance cues for re-identifying a person. Due to this, homogeneous Re-ID can be limited in surveillance and may not capture reasonable appearance information under poor illumination conditions.

On the other hand, heterogeneous Re-ID [5, 28–30] assumes that the query and gallery data comes from different modality. Though it allows deployment under real world scenarios where cameras capture under different spectrums or the probe is a natural language description,

heterogeneous Re-ID task scales up the challenges during matching the person or probe against the gallery.

In this dissertation, we address both the categories of the Re-ID problem. We propose video-video and RGB-RGB Re-ID under homogeneous settings and RGB-IR and text-image Re-ID in heterogeneous settings. Figure 1.2 shows the examples for both categories with visual examples.

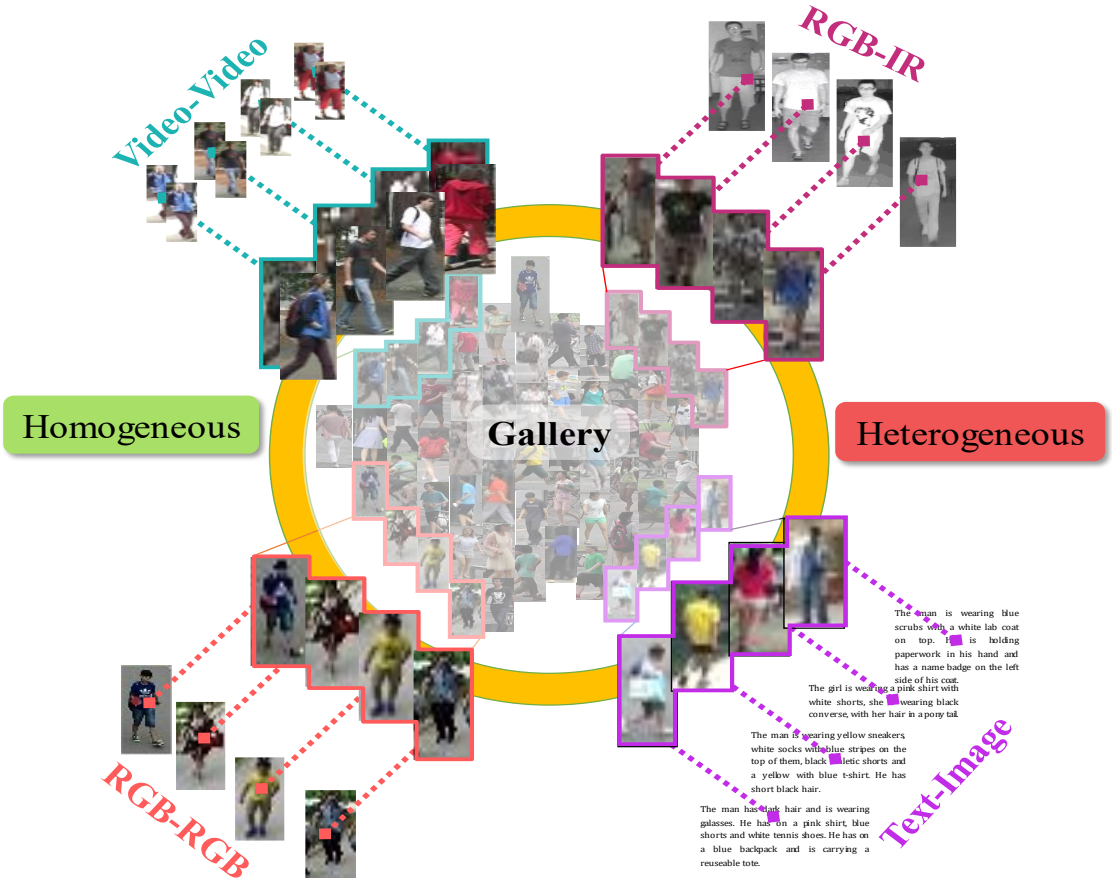


Figure 1.2: *Homogeneous and Heterogeneous Re-ID System*

1.3 Challenges

Re-ID is a difficult task as it suffers from severe challenges [12, 14, 31–33]. These challenges arise due to the fact that the probe and gallery samples come from different settings or modalities. Figure 1.3 illustrates the various challenges in both categories with visual examples.

1.3.1 Homogeneous Re-ID

- **Re-ID Challenges:** Re-ID suffers from the challenges of pose variation, illumination variation, occlusion, and low resolution. Learning a deep model that can generalise well

against the various challenges is a hard problem. Further, the non-availability of large labeled datasets as well as availability of only few samples per person per camera poses a great challenge in training deep networks and often lead to overfitting. Different from generic recognition problems, Re-ID also suffers from the large intra-class diversities and inter-class differences, that is different persons can look alike across camera views, and the same individual may look different under different camera views.

- **Video based Re-ID:** In addition to the above Re-ID challenges, in case of videos, the problem becomes even more challenging due to the inherent dynamic nature of videos and motion blur.

1.3.2 Heterogeneous Re-ID

The main drawback of the homogeneous Re-ID is that it assumes ideal and fixed environmental conditions like good visible lighting and limits the applicability of person Re-ID in practical surveillance applications. In addition to the homogeneous challenges, it has many more challenges as follows.

- **RGB-IR Re-ID:** Though the wide applicability of heterogeneous settings where devices have both visible and infrared imaging capability is encouraging, it is more complicated compared to homogeneous Re-ID. In addition to RGB-RGB Re-ID challenges, the spectrum challenge also needs to be addressed. In particular, IR images are void of color and texture. Thus learning robust representation which can match RGB and IR images is very hard.

- **Text-Image Re-ID:** Under heterogeneous settings, we also address the problem where only a textual description is available for a person. The completely different modality of image and text further adds to the existing challenges in Re-ID. Additionally, the text modality also requires investigation of other feature representation models such as Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTMs).

In the following section, we briefly discuss our contributions.

1.4 Research Contributions

In this work, we propose homogeneous (video-video and RGB-RGB) and heterogeneous (RGB-IR and text-image) Re-ID systems. The contributions are summarized as follows:

- **Video-Video Re-ID:** We propose a novel shallow end-to-end network for video-video matching [25, 34, 35]. To learn rich representations from multi-scale visually attentive regions, we use CNN and spatial pyramid pooling (SPP) [36] network. Since discriminative frames may appear anywhere in the sequence, we exploit spatio-temporal fusion of features by adding spatially pooled features of CNN with temporally pooled features of RNN. In order to train the network, we use a multi-loss function. Our model requires

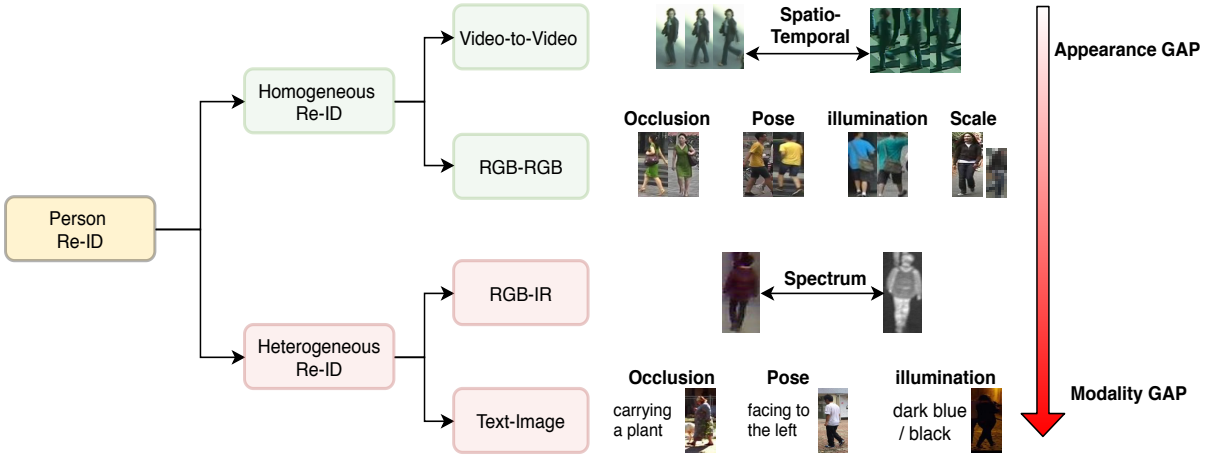


Figure 1.3: *Challenges of Person Re-ID. Green boxes shows the challenges in homogeneous Re-ID models and red boxes shows the challenges in heterogeneous Re-ID with examples. Color coded arrow indicates the challenges including the appearance gap to modality gap.*

only approximately 15% of the parameters compared to state of art deep models such as LuNet [37]. We also contribute a large novel dataset of air borne videos for mobile platform video person re-identification, named as DJI01. It is taken under various challenging conditions like view variation, occlusion, illumination changes, people with similar clothes, same people at different days, scale variation and pose variation.

- **RGB-RGB Re-ID:** We propose a robust multi-resolution encoder decoder network to learn features invariant to pose, occlusion and illumination [25, 38]. The additional generative task makes the learned feature representation richer and thus more generalisable to the test data. Since the images generated by the decoder are from multiple scales, the learned features are also scale invariant to a large extent. We also show qualitative analysis to verify the claim. We demonstrate a new mechanism of test data augmentation using the reconstructed images. Such augmentation can further enhance accuracy in case of single query setting. The increase in accuracy can be attributed to the fact that reconstructed images are invariant to multiple challenges, thereby increasing the feature robustness.

- **RGB-IR Re-ID:** In addition to the homogeneous Re-ID problems, we address the problem of heterogeneous RGB-IR Re-ID by exploiting the Spectrum-Disentangled representation Learning (SDL) [30]. To the best of our knowledge, in the person Re-ID domain, it is the first effort to design a network with disentanglement loss to filter out the spectrum information. To achieve the goal of spectrum invariant features, we design two branches in our network, named as spectrum dispelling and spectrum distilling branch. The first branch is to keep the useful cues specific to identity. On the other branch, we apply an identity-dispeller loss to fool the identity classifier so that it primarily learns spectrum related information.

- **Text-Image Re-ID:** We address the problem of fine-grained description based Person Re-identification by exploiting the Hierarchical Attention Alignment Network (HA2-Net). We introduce a novel strategy of Term Frequency-Inverse document Frequency

thresholding to extract the most salient tokens to alleviate the challenge of *text complexity*. Second, a hierarchical attention alignment is proposed to determine the potential relations at different levels: word-patch level, phrase-patch level, and sentence-image level for addressing *alignment uncertainty*. The network is optimized via joint weighted hierarchical attention loss and cross-modal classification and matching losses in an end-to-end manner.

1.5 Dissertation Organization

The rest of the work is structured as follows.

- Chapter 2 discusses the literature work for homogeneous and heterogeneous based Re-ID system.

- In chapter 3, we present the proposed algorithm for video-video Re-ID system. Further, we introduce a new air borne video dataset with challenging situations.

- In chapter 4, we present our proposed approach for homogeneous image based Re-ID, and analyze each component of the network. First, a deep reconstruction Re-ID network using hybrid sampling strategy is discussed. Further, a new mechanism of test data augmentation is followed by experiments on publicly available datasets to demonstrate the state-of-the-art performance.

- Chapter 5 describes our proposed approach for heterogeneous image based Re-ID. This chapter addresses the problem of cross-modality RGB-IR Re-ID by exploiting the Spectrum-Disentangled representation Learning (SDL). Further, effectiveness of our proposed method on challenging RGB-IR Re-ID datasets is described.

- Chapter 6 extends the Re-ID problem into more broader view from person to other categories. The problem of fine-grained text-image matching is addressed. First, a novel strategy of Term Frequency-Inverse document Frequency thresholding is described. The hierarchical attention alignment approach is then detailed followed by experimental results on standard challenging datasets.

- Finally, chapter 7 concludes the thesis by summarizing the contributions and discusses the future research work on Re-ID.

Chapter 2

Literature Survey

In this chapter, we provide an overview of homogeneous and heterogeneous techniques. In section 2.1, we discuss homogeneous based Re-ID which covers algorithms for video-video and RGB-RGB. Further, in section 2.2, we discuss closely related work which deals with the RGB-IR and text-image work of heterogenous Re-ID.

2.1 Homogeneous based Re-ID

These methods can be categorized into two different classes: video-video and RGB-RGB matching models.

2.1.1 Video based Person Re-ID

Video based person Re-ID is a challenging and important task in surveillance based applications. Several approaches exploit spatio-temporal information for video-based person Re-ID [21, 39–41]. Mclaughlin *et al.* [21] use RGB frames and optical flow between consecutive frames to construct the spatio-temporal features. However, its single stream convolutional neural network (CNN) takes partial advantage of rich temporal information. In other tasks such as action recognition, two stream CNN architecture has been shown to give better performance [42]. Video is split into two individual streams, i.e. sequence of still images and optical flow vectors, to separately learn better representative features. The information from the two streams is then fused at certain intermediate layers. Shallow models like RCN [21] and ASTPN [43] perform well on small datasets such as PRID-2011 [1], iLIDS-VID [3]. However, the performance does not generalize to large datasets such as MARS [2].

Aerial Video based Re-ID Models

Schumann *et al.* [44] propose to use colour and texture features to recognize individual persons in aerial video data and features are weighted based on their correlation to operator feedback in order to find possible matches to a query person track. However, the dataset captured with moving platform requires view specific discriminative training for obtaining good performance. This is because of the view variation in the dataset, where view variation is defined as continuously varying view angle of the camera compared to traditional static cameras [45]. The diffi-

culty to collect large view specific data limits the learning ability of a model. Layne *et al.* [45] introduce a mobile platform database and test with SVM and metric learning techniques. However, it is captured at very low altitude and does not have large variance per identity. In [46, 47] authors exploits person Re-ID in aerial images. Good surveys of aerial based models for video-video Re-ID can be found in [3, 44, 45, 48, 49].

Handcrafted feature based Models

Wang *et al.* [3] propose a model based upon optical flow and spatio-temporal features which select and match discriminative video fragments from unregulated pairs of image sequences of people. Other popular spatio-temporal features are 3D SIFT [48] and 3D HOG [50]. Jinjie *et al.* [49] use 3D HOG descriptor for extraction of the space-time features. 3D HOG feature contains spatial gradient and temporal dynamic information. In addition, they use color histograms and local binary pattern (LBP) features to describe a person's appearance in each frame. Similarly, color features have been explored by Liu *et al.* [51]. However, low level color features are not discriminative enough to distinguish persons and are sensitive to changes in appearance due to lighting and view changes. Simmonet *et al.* [52] apply dynamic time warping (DTW), which is a popular sequence matching method widely used for action recognition for person re-identification task. Nixon and Carter [53] study gait pattern for video based person re-identification. However, it requires accurate silhouette extraction which is still an open problem. All these works use either multiple images or a selected fragment of a sequence to extract features. However, the rich spatial and temporal information may not be fully exploited which can degrade the performance.

Deep Learning Models

Over the past decade, deep learning methods [54, 55] have shown a significant improvement over hand-crafted features. They encode reliable features and corresponding similarity value for a pair or triplet of images or videos. In last few years, deep learning has shown incredible performance in many tasks such as action re-localization, person Re-ID and search. Spatio-temporal action relocalization has received significant attention in the computer vision communities [56, 57] due to its vast applications such as public security, event recognition and video retrieval. There are some publicly available datasets such as UCF-Sports [58] and UCF101 [59], which have made great contribution to improve the performance for the task of action relocalization. Based on these benchmarks, there are a few promising deep model based methods, including two streams framework [42], C3D [60], and Artnet [61] for action recognition. These methods mainly try to extract different vision cues, such as short video clips [60], motion information [42] and long-range video clips [62]. Deep learning architectures based on recurrent neural network (RNN) [21] and long short-term memory (LSTM) [39–41] models have been explored for person Re-ID. McLaughlin *et al.* [21] focus on color appearance and optical flow, where the network jointly learns feature representation and similarity metric. RNN efficiently encodes the temporal information, but there is limitation with learning of the long duration sequences of the inputs [21]. Graves [39] use an LSTM model for learning long duration dependencies through the use of memory cell units. Varior *et al.* [40] propose a siamese LSTM architecture that can process image regions sequentially and enhance the discrimina-

tive capability of local feature representation by leveraging contextual information. Feedback connections and internal gating mechanism of the LSTM cells enable the model to memorize spatial dependencies and selectively propagate relevant contextual information through the network. Wu *et al.* [41] propose a framework which combines time series modeling and metric learning to jointly learn relevant features and similarity measure between time sequences of person. However, LSTM requires large training dataset due to large number of parameters for achieving good generalization. Yan *et al.* [63] investigate a recurrent feature aggregation network (RFA-Net) with LSTM Layer. This network uses a set of hand-crafted frame level features extracted from each frame in the input sequence and produce sequence-level features. However, these features may exhibit limited discriminative nature. Dai *et al.* [19] propose a temporal residual learning module which is equipped with two bi-directional LSTMs (BiLSTMs) to simultaneously learn the generic and specific features of a video sequence. Xu *et al.* [43] propose ASTPN takes the advantage of attention mechanism to extract features from informative frames. Subramaniam *et al.* [64] propose an attention driven approach that exploits a common set of salient features across multiple frames of a video. Zheng *et al.* [2] use Alexnet to extract features and metric learning to compute the similarity. Zhou *et al.* [65] propose the temporal attention model (TAM) to focus on discriminative frames. The TAM is jointly learned with the spatial recurrent model (SRM) to integrate the surrounding information at different spatial locations for better similarity evaluation. Zhong *et al.* [66] propose a re-ranking method in which given an input, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector, which is used for re-ranking under the Jaccard distance. Chung *et al.* [67] use two stream siamese network to learn spatio-temporal features separately for person re-identification. Yu *et al.* [68] explores different streams to learn different aspects of feature maps for attentive spatio-temporal fusion of video, and then merges them together to study union features. Li *et al.* [69] propose a DCF model, in which Multi-Scale Context-Aware Network (MSCAN) is designed to learn powerful features over full body and body parts, which can well capture the local context knowledge by stacking multi-scale convolutions in each layer. Li *et al.* [70] propose SATPP model to leverage the rich visual-temporal cues for feature learning.

Loss Functions

In addition to exploration of different features and architectures, various loss functions have also been investigated. Siamese and triplet loss have been exploited in [21] and [37] respectively. The triplet loss reduces the intra-class variation and increases the inter-class variation. However, Hermans *et al.* [37] does not fully take the label information of the data into account as their goal is to pull the instances of the same person closer and push the instances of different persons apart. Further the authors highlight the fact that triplet loss based training for pre-trained models as well as models trained from scratch can achieve state-of-art performance. The two models, TriNet and LuNet, are derived from pre-trained ResNet-50 and ResNet-v2 architecture respectively. Authors highlight that due to the efficacy of triplet loss even the LuNet model which is trained from scratch achieves a comparable performance to TriNet model. We observe that shallow networks work well in case of smaller datasets like PRID-2011 and iLIDS-VID. However, in case of larger dataset like MARS there is large drop in accuracy. To deal with larger datasets, one can use deep pre-trained models. Algorithms in [2, 37, 69] uses AlexNet,

GoogleNet and ResNet and achieve state-of-art performance on MARS dataset. In addition, [37] uses LuNet model which is trained from scratch and show comparable accuracy with the use of triplet loss. However, LuNet uses 5M number of parameters whereas our shallow network uses only 0.75M parameters with triplet loss and achieves comparable accuracy to that of LuNet. Recently, attention models have shown good performance in various tasks. Haque *et al.* [71] exploit an attention model for the depth data to learn a specific local region.

In this thesis, we use attention based deep learning models to exploit the spatio-temporal fusion features for video-video matching. In next section, we discuss state-of-the-art algorithms for the RGB-RGB based work.

2.1.2 RGB-RGB based Person Re-ID

Prior work on still images can broadly be classified into two main categories. The first one is feature extraction and learning discriminative models corresponding to those features [72–75], while the other is learning based technique that learn the features in an end-to-end manner. Traditional approaches typically employ hand-crafted features and a metric learning model. They either focus on designing distinctive and view-invariant feature representations [73, 76–78], or learning optimal similarity/distance metrics [66, 72, 74]. Recently, deep learning methods [13, 22, 55, 79–82] demonstrated great success and have the advantage of learning the optimal feature representation and matching metrics jointly in an end-to-end manner through various losses [37, 79, 80, 83–85].

Deep Learning Models

Learning based methods encode the reliable features and corresponding similarity values for a pair or triplet of images [79, 83, 84]. It removes the need of extracting hand-crafted features. With the great success of CNN features, a lot of recent Re-ID approaches [22, 55, 81, 86–88] are designed based on CNN structure. These methods can be categorized into four different classes-Attention, Part-based, Global feature extraction and Multi-resolution models.

Attention Models: In order to learn a focused foreground object, attention masks are learned [89–91]. Liu *et al.* [89] feeds multi-level attention maps to different feature layers to capture multiple attentions from low-level to semantic-level. Liu *et al.* [90] propose Contextual-Attentional Attribute Appearance network (CA^3Net) to jointly learn visual attention on attributes and appearance representation by multi-task learning and attribute recognition. Li *et al.* [91] introduce Harmonious attention network (HA-CNN) to process multiple focused regions in different branches and merge with the global features in the final layer for joint learning of hard regional attention and soft pixel attention. Si *et al.* [92] propose Dual ATtention Matching network (DuATM) via dual attention block which simultaneously performs feature refinement and feature-pair alignment to learn context-aware feature sequences. Zheng *et al.* [93] exploits Consistent Attentive Siamese Network (CASN) that integrates attention consistency modeling and Siamese representation learning in a joint learning framework to learn identity aware invariant representations.

Part-based Models: One of the major challenges for person Re-ID is the diversity of human

pose. To obtain a pose invariant representation, several part-based works [22, 55, 88, 94, 95] are proposed. AlignedReID [22] explores local features and introduces a feature matching method to align different body parts. We *et al.* [55] explores Global-Local Alignment Descriptor (GLAD) to explicitly leverages the local and global cues in human body using part extraction from regions. Beyond parts model (PCB) [88] uses part based models with uniform partition strategy and employ information from different parts of the image. However, partition strategy still remains a challenge. Qian *et al.* [94] propose a Pose-normalization Generative Adversarial Network (PN-GAN) which allows a deep person image generation model for synthesizing realistic person images conditional on pose to enhance the scalability of the model. However, the global information, a very significant cue for Re-ID, is completely ignored in this model. Zheng *et al.* [95] propose the pose invariant embedding (PIE) as pedestrian descriptor via PoseBox fusion network to address the pedestrian misalignment problem.

Global Feature Extraction Models: Xiao *et al.* [86] propose a domain guided dropout algorithm to improve the feature learning procedure by joint training of multiple Re-ID datasets. DarkRank [96] shows that a powerful teacher model can significantly help the training of a smaller and faster student network for Re-ID. Huang *et al.* [81] determine the discriminative regions in the training data and adversarially occlude a certain percentage of this area. These samples are then used for training data augmentation, which essentially helps the network to come out of the local minima during training. Shen *et al.* [97] propose a Similarity-Guided Graph Neural Network (SGGNN), which incorporates graph computation and inter-gallery-image relations in the training stage to enhance feature learning process in an end-to-end manner. Subramaniam *et al.* [98] propose light-weight fused siamese network by exploiting spatial correspondence structures via the patch comparison layer to handle the challenges of illumination variations, partial occlusions and viewpoint invariance.

Multi-resolution Models: All the above discussed methods typically consider only one input image resolution. Due to this, it drops the potentially useful information of other resolutions. Few works like multi-scale TriNet (MS-TriNet) [99], multi-scale deep learning model (MuDeep) [100], and deep pyramid feature learning (DPFL) [101] exploits information from multiple scales in Re-ID. In MS-Trinet, Liu *et al.* [99] propose a multi-scale network and combine features from three different scales of low resolution by a hard embedding layer and learn a multi-branch CNN model. In MuDeep, Fu *et al.* [100] propose a verification network to exploit the multi-scale features for Re-ID. DPFL [101] consists of m scale specific branches each for learning one input image scale in the pyramid, and an additional scale-fusion branch for learning complementary combination of multi-scale features. In MS-Trinet [99] and DPFL [101], authors utilize multi-scale information at the input end whereas Fu *et al.* [100] uses MuDeep and takes advantage from multiple scales at the convolution layers.

After feature extraction, various metric learning methods [66, 102] are used to find out the similarity between a pair of images or videos. Most approaches choose the Euclidean distance to compute a similarity score for a ranking or retrieval task [7, 37]. In particular, [66] propose a re-ranking (RK) method with k-reciprocal encoding, which combines the original distance and Jaccard distance to improve Re-ID accuracy.

Several approaches define a new loss function for person Re-ID [37, 83, 84, 86]. Cheng *et al.* [84] propose an improved triplet loss by introducing another pull term into the triplet loss [37, 83], penalizing large distances between positive embeddings. Chen *et al.* [85] design

a quadruplet loss, which employs four samples and adds another pull term for the distance between negative pairs, which can lead to a model with a smaller intra-class variation and larger inter-class variation. However, generation of samples of triplets or quadruplets still remain a challenge.

Sampling Strategy

Since generation of relevant triplets is a challenging task, in triplet network (TriNet) [37], authors propose a batch hard sampling strategy and show significant performance improvement. This strategy is also applied in [22, 100]. However, it is not an optimal approach. Fan *et al.* [80] present a balanced sampling strategy to remove the bias in number of available samples per person. Here, they use sampling without replacement when there are sufficient samples for a person, and with replacement when the samples are less.

In this thesis, we propose a deep reconstruction method to learn generalizable features and address all the conventional Re-ID challenges simultaneously. We discuss the state-of-the-art algorithms for the RGB-IR based work in the next section.

2.2 Heterogenous based Person Re-ID

We cover two major types of heterogenous Re-ID methods which includes RGB-IR and text-image based models.

2.2.1 RGB-IR Re-ID

RGB-IR Re-ID has received far less attention compared to RGB-RGB Re-ID problem. In addition to the challenges present in RGB-RGB Re-ID, the spectrum variation poses a great challenge in RGB-IR case. In one of the first works, Wu *et al.* [5] proposed a deep zero-padding framework for shared feature learning under two different modalities. The deep zero-padding promotes domain-specific learning efficiently. The domain specific nodes mainly occur in shallower layers and the deep layers are shared. In particular, the domain specific nodes are responsible for bridging the gap between the domains. [5] has also been used in other modalities like RGB-Depth in [28]. Ye *et al.* [29] introduced a two-step framework for feature learning and metric learning. Feature learning involves training using identity loss and contrastive loss. The metric learning comprises of learning the intra-modality transformation matrices for better clustering of samples from same person, and the metric to push inter modal samples of different persons farther apart. Recently, Ye *et al.* [4] proposed an end-to-end dual-path network to learn common feature representations and showed significant improvement.

Feature Disentanglement

Disentangling approaches have been used in different computer vision tasks. Some works proposed to disentangle the representations in face tasks such as pose invariant recognition [103]

and identity-preserving image editing [104]. These methods encode feature vector based on attribute supervision. However, if new attribute gets added then it needs re-training. Liu *et al.* [105] proposed to learn disentangled face identification in addition to face features. The complete input information is encoded in identity distilled and dispelled features using an auto-encoder trained adversarialy via MSE loss. MSE being a pixelwise loss yields poor reconstruction results when there is a huge pose difference. Thus a robust encoding of the image in terms of identity distilled and dispelled features may not be realized. Bousmalis *et al.* [106] propose an unsupervised domain adaptation technique wherein shared and private sub-spaces are explicitly modeled to obtain better domain invariant representations. DSN learns these shared and private representations using labeled source and unlabeled target datasets which have the same classes. A scale-invariant mean squared error loss is applied to reproduce the overall shape of the objects along with classification, MMD [107] and DANN [108] losses. However, a reconstruction loss based on pixels may not be a good choice in learning better Re-ID features. This is because, the images from different cameras are very distinct in pose, illumination, scale and spectrum. Disentangled representation has also been exploited in [109–112].

Adversarial Learning

Recent Re-ID methods leverage GANs to fill the domain gap between multi-domain and multi-modal datasets. Ganin *et al.* [108] exploits the idea of augmenting a network with discriminator to enforce hidden representation invariance with the goal of domain adaptation. Metzen *et al.* [113] used this kind of architecture in the context of detection. The discriminator, trained separately from the classifier, is merely used as detector of adversarial attacks. Image-to-image translation has been widely studied using GANs [114, 115]. Based on the translation of RGB image to thermal images, Kniaz *et al.* [116] train a ThermalGAN to obtain multiple thermal images which are then used for matching against a thermal image gallery. Dai *et al.* [117] proposed *cm*GAN to learn discriminative feature representation from different modalities. However, model with adversarial part is hard to train and the convergence rate is less promising.

In this thesis, we address the RGB-IR based heterogeneous problem via fast disentanglement method. Our proposed method does not use adversarial learning techniques, converges quickly and obtains strong discriminative identity-related features while disentangling the spectrum information. In the following, we discuss another emerging area of text-image Re-ID.

2.2.2 Text-Image based Person Re-ID

Text-image matching has gained a lot of attention in recent years due to its wide applications, such as product searches in web-pages and person retrieval in the video surveillance system. Different from existing cross-modal matching methods [18, 30] where both query and gallery are image-based, the text-image matching utilizes text descriptions for query or gallery. It is of significant practical usage as a query image is difficult to obtain in many cases, and on the other hand, textual descriptions can be easily accessible.

Most of the existing text-image matching methods [118, 119] try to learn a joint mapping that projects the corresponding text-image pair to a shared subspace. However, the learned global representation may also encode irrelevant details such as complex background in images

and often disregard the complexity of the text domain. Since it is a challenging task to fill the large domain gap between text and image data, such global representations may not be efficient. In order to overcome the issues of global representations, some of the existing fine-grained matching [10, 120, 121] techniques learn the relation between all possible pairs of words and regions without attending the important words and regions. Moreover, this approach disregards the fact that there are patches or words which do not have any correspondence.

Global Feature Embedding Based Methods

Frome *et al.* [122] designed a Visual Semantic Embedding (VSE) framework to obtain the global features of input images and text in common feature space by using a ranking loss. Faghri *et al.* [123] proposed the VSE++ model and penalized the existing model via the hardest negative examples and improved the alignment of image-sentence. Different types of loss functions such as bi-rank [124], Histogram [125] and N-pair [126] loss are also designed to match global embeddings. However, the performance is largely dependent on batch size. Bi-rank [124] produces comparative matching accuracies with a larger batch size. The N-pair loss proposed an efficient batch construction to address the slow convergence of triplet loss. [118] introduced cross-modal loss functions for projection matching (CMPM) and classification (CMPC) to learn discriminative text-image representations. Wang *et al.* [119] proposed a mutually connected classification loss (MCCL) to promote the discriminative feature embedding. However, feature embedding based strategy ignores the importance of fine-grained visual-semantic similarities of text-image pairs. On the other hand, the learned global representation suffers from irrelevant information generated by background and other uninformative visual regions.

Attention Based Methods

Li *et al.* [10] proposed the first large-scale person description dataset CUHK-PEDES, which contains person images with detailed natural language annotations. They also provided an attention based model, Recurrent Neural Network with Gated Neural Attention (GNA-RNN) with unit-level attentions and word-level gates to determine the cross-modal affinity. Further, Li *et al.* [127] proposed an identity-aware two-stage framework based on the co-attention mechanism for text-image matching. However, the attention is paid to one single direction of LSTMs to get word representations which may incur some noise because a hidden state contains a complex semantic mixture of the current word and previous words. Huang *et al.* [128] proposed a selective multi-modal Long Short Term Memory network (smLSTM) which selectively attends to a pair of instances of image and sentence at each time step for text-image matching. Chen *et al.* [129] improved visual representations through local and global relationships between text and image. The global relationship is guided via identity labels, and the local association focuses on improving the visual representations by phrase reconstruction. However, they simply aggregate the similarity of all possible pairs of regions and words without attending differentially to more and less important words or regions. Different from the above methods where architectures are based on CNN-RNN, Dual Path [130] employs CNN for textual feature learning. Hu *et al.* [120] proposed a relation-wise dual attention network (RDAN) for text-image matching. Beyond the fundamental text-image matching, the fine-grained problem is the ma-

jor difficulty in distinguishing different classes that need to be carefully addressed. Therefore, different from the methods above, we propose to attend the most salient text and match visual human parts, and further propose an attention alignment model to carry out more accurate fine-grained text-image matching.

The subsequent chapters details the proposed methods for problems discussed in this chapter and outline the comprehensive experiments and results.

Chapter 3

Video based Person Re-ID

Video can naturally be attributed to spatial and temporal cues. The spatial part carries information about scenes and appearances of the persons–like color of the clothes, person' height and shape, while the temporal part encodes the walk pattern of the person, which is complementary to the spatial part. Thus, learning the representations from videos can lead to better results. In this chapter, we discuss our novel architecture named as CARF-NET for video based person Re-ID in section 3.1. Further, we discuss our novel DJI01 dataset for air-borne videos with challenging situations in section 3.2. Then, we elaborately discuss the experiments and results in the further sections.

3.1 Proposed Algorithm

In this section, we propose a novel two stream triplet network for video based person Re-ID. The input consists of color channels and optical flow vectors, where optical flow is computed using Lucas-Kanade algorithm [131]. Color channels encode details of a person's appearance and clothing, whereas optical flow encodes the temporal information. We first use a CNN module to extract the spatial information. CNN is followed by a spatial pyramid network which exploits visually attentive regions in the inputs. In order to encode the temporal information, we use RNNs. In addition, we perform fusion of spatial features with recurrent layer features across time steps to form a discriminative video-level representation. We train the network using triplet softmax loss. This approach preserves the spatial information along with temporal information and also learns the most contributive regions towards Re-ID. A diagram of our proposed architecture is shown in Fig. 3.1.

3.1.1 CNN with Spatial Pyramid Pooling for Visual Attention

The two streams in the architecture use spatial network (CNN) for learning feature representations from raw frames and optical flow respectively. Let an input video sequence be denoted as $s_t, \forall\ t = 1, 2, ..., T$, where T is the length of the sequence. Each layer of the convolutional network performs the following operation:

$$C_{l+1} = tanh(maxpool(conv(C_l))), \tag{3.1}$$

where l denotes a layer and C_0 denotes the raw video frame or optical flow, and in deeper layers the input is the output feature map from the previous layer of the CNN. To avoid complexity

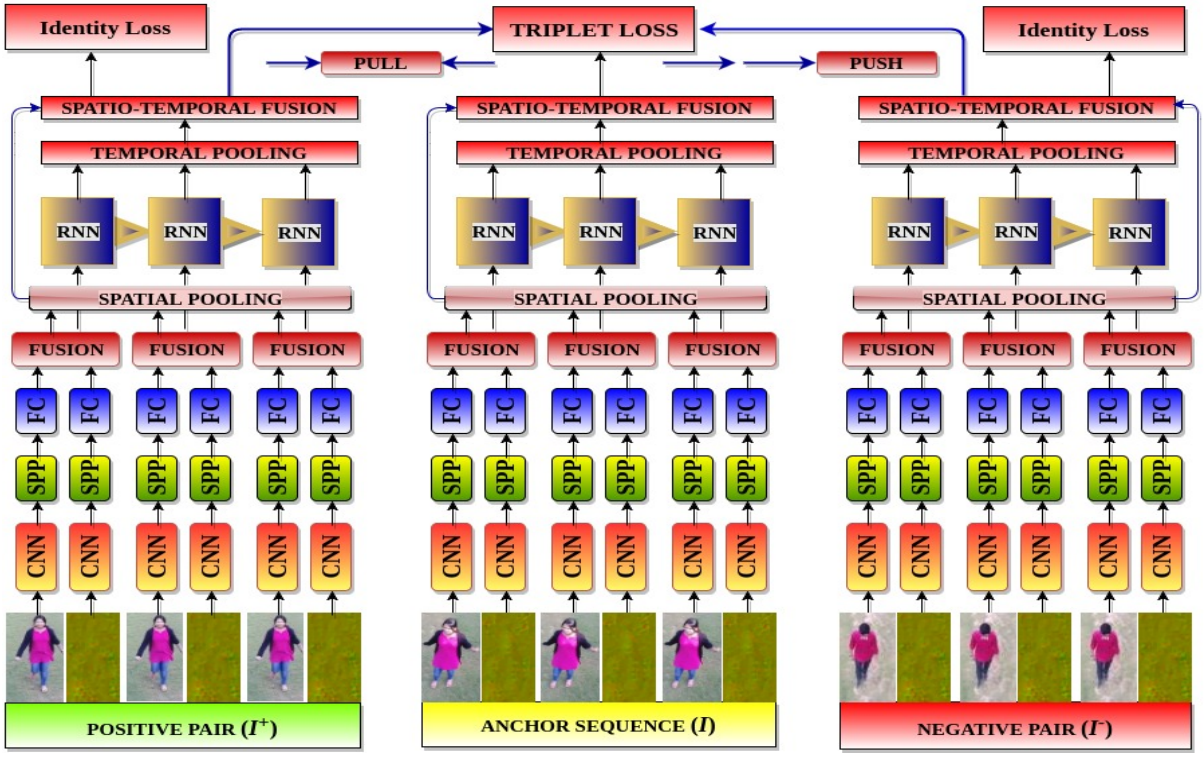


Figure 3.1: *Proposed Architecture CARF-Net. The input to the two stream spatial network (CNN) comprises of color image and optical flow. SPP refers to the spatial pyramid pooling. FC denotes 3 fully connected layers. Outputs of the last fully connected layer from the two streams are fused in FUSION layer and the resultant feature is given as an input to RNN. Simultaneously, the resultant features are also pooled together in SPATIAL POOLING layer. The output of RNN layer is pooled in TEMPORAL POOLING layer. Further, outputs of SPATIAL POOLING layer and TEMPORAL POOLING layer are fused in SPATIO-TEMPORAL FUSION layer.*

and high computation, we use an efficient spatial network as in [21]. The spatial network (CNN) contains 3 convolutional layers and 2 max-pooling layers with a non-linear *tanh* layer. Further, we use an attention mechanism [36], which exploits attentive component to select discriminative local spatial regions from each raw video frame or optical flow. This is necessary as only sub-regions of the entire frame contribute towards re-identification. Thus, selecting only those regions can enhance the accuracy. We use spatial pyramid pooling (SPP) [36] as an intermediate layer over last convolution maps to concentrate over important regions in spatial dimension. We use SPP because it generates multi-level spatial representations which exploits the information from various scales. Let the feature maps set be $F = \{F_1, F_2, ..., F_T\}$, obtained from the convolutional network. Each $F_i \in R^{c \times w \times h}$ is then fed into spatial pooling layer to get image-level representation I_i, where c is the number of kernels and, w and h are kernel' width and height respectively. SPP layer has spatial bins to generate multi-level spatial representations. The dimensions of spatial bins are $(B_w{}^j, B_h{}^j | j = 1, 2, ..., n)$, the window size $win_j = \left(\left\lceil \frac{w}{B_w{}^j} \right\rceil, \left\lceil \frac{h}{B_h{}^j} \right\rceil \right)$ and pooling stride $str_j = \left(\left\lfloor \frac{w}{B_w{}^j} \right\rfloor, \left\lfloor \frac{w}{B_h{}^j} \right\rfloor \right)$ for j^{th} spatial bin. Finally, the resultant vector I_i is as follows:

$$v_{i,j} = f_{reshape}\left(f_{pool}(F_i; win_j, str_j)\right) \tag{3.2}$$

$$I_i = v_{i,1} \oplus v_{i,2} \oplus .... \oplus v_{i,n} \tag{3.3}$$

where f_{pool} denotes the max pooling with window size win and stride str, $f_{reshape}$ represents the reshape operation to reshape matrix into a vector and $\oplus$ denotes the vector connection operation. Multi-level spatial representations generated from spatial bins are then combined into a fixed-length image-level representation. These representations involve the exact position of a person and multi-scale spatial information. We use a 4-level pyramid pooling for local spatial attention after the last convolution layer. The output of SPP layer is fed to the first fully connected (FC) layer. We use 3 FC layers. The output of the third FC layer is a 128 dimensional feature vector. Similarly, feature vector is generated from the second stream. We fuse the two feature vectors from two streams to obtain a single representation. We explain the fusion operation in Section 3.1.3. The fusion is represented by the component 'FUSION' in Figure 4.1. Further, let $r = r_t \in R^{128\times 1} | t = 1, 2, ..., T$ be a sequential representation of the fused features from the two streams. In order to encode temporal information, we then pass r to the RNNs which we explain next.

3.1.2 RNN

Given an input sequence $r_t, \forall\ t = 1, 2, ..., T$, the output of the RNN is o_t which is the final feature vector obtained at time t. The RNN equation is given by,

$$o_t = W_r r_t + W_h h_{t-1} \tag{3.4}$$

$$h_t = tanh(o_t) \tag{3.5}$$

Here, $o_t \in R^{q\times 1}$ is the q-dimensional output of RNN at time-step t, $h_{t-1} \in R^{q\times 1}$ contains the information on the RNNs state at the previous time-step, W_r and W_h are the respective weights for r_t and h_{t-1}. We add a temporal pooling layer after RNN to capture long-term information present in the whole sequence. Although RNNs are able to encode the temporal information, they are biased to current information [21]. However, discriminative frames may appear anywhere in the sequence. In order to preserve the information from discriminative frames, we apply the temporal pooling. We use an average-pooling over the temporal dimension to produce a single feature vector, which we explain next.

3.1.3 Fusion

In this section, we discuss fusion, spatial pooling, temporal pooling and spatio-temporal fusion layers present in the architecture. Let f_{FCS1_t} and f_{FCS2_t} be the respective outputs of FC layers from streams 1 and 2 at a given time instant t, respectively. Then the fusion of these features is given by,

$$f_{FUS_t} = \frac{1}{2}(f_{FCS1_t} + f_{FCS2_t}) \tag{3.6}$$

where f_{FUS_t} is the same as r_t.

Spatial Pooling: The inputs to spatial pooling layer are f_{FUS_t}. The pooling equation is given by,

$$f_{SP} = \frac{1}{T}\sum_{t=1}^{T} f_{FUS_t} \tag{3.7}$$

where T is the length of the sequence or time-steps.

Temporal pooling: Let f_{Temp} denote the motion information averaged over the whole input sequence, then

$$f_{Temp} = \frac{1}{T}\sum_{t=1}^{T} o_t \tag{3.8}$$

Spatio-Temporal Fusion: We observe that the spatial information along with the temporal information can be a better representative for re-identification. Towards this, we pool the spatial and temporal features and this operation is given as in [132]:

$$F_{SPT} = f_{SP} + f_{Temp} \tag{3.9}$$

In our experiments we find that there is 2% to 4% increase in the performance due to spatio-temporal fusion.

3.1.4 Training Objective

The triplet network consists of three sub-networks with identical weights. In order to train the triplet network to perform re-identification, our objective is to show similar and dissimilar input pairs, and it must learn to map those inputs to a feature space where similar inputs are close and dissimilar inputs are separated by a margin. We use a multi-loss function including triplet loss [133] and identity loss. We expect that the features of the positive samples ${I_n}^+$ are more similar to anchor sequence I_n than the features of the negative sample. Here, I_n can be represented by a feature, in our case we use F_{SPT}. Let α be the margin between positive and negative pairs to enhance the discriminative ability of learned features. Thus, we have:

$$\|I_n - {I_n}^+\|_2^2 + \alpha < \|I_n - {I_n}^-\|_2^2 \tag{3.10}$$

The proposed network can be trained end-to-end using back-propagation and the loss function for N triplets is given by:

$$L_{triplet} = \frac{1}{N}\sum_{n=1}^{N}\left[\|I_n - {I_n}^+\|_2^2 - \|I_n - {I_n}^-\|_2^2 + \alpha\right]_+ \tag{3.11}$$

where $[.]_+$ is the hinge function.

Given the sequence feature vector I, we can determine the identity of the person in the sequence using the standard softmax function. Let Γ be the total number of identities, z is the predicted identity for the input person, and $S \in R^{M\times\Gamma}$ is the weight matrix used in the softmax

Figure 3.2: *First column refers to the anchor. Second column shows positive samples corresponding to the anchor. Third column shows the negative samples.*

ALGORITHM 1: Training of the Network

Input: An anchor sequence: $A = (x_u^a)_{u=1}^N$; a positive set: $P = (x_u^p)_{u=1}^N$ and a negative set: $N = (x_u^n)_{u=1}^N$ (shown in Figure 3.2) with optical flow of video sequence s_t, $t = 1, 2, ...T$ where T is the sequence length and N is the number of triplets.
Output: The network Parameters W
Initialize W with small random values;
repeat
  Randomly sample k, $k = 1, 2, ...K$ where K is the number of identities;
  for $n = 1, 2, ..N$ **do**
    Randomly sample three video sequence for each selected k^{th} id: first from A, second from P and the third from N, to form training batch;
    Run forward propagation for each sequence;
    Extract Spatio-Temporal Fusion features: F_{SPT} using Equation (3.9);
    Calculate gradient using back-propagation;
    Update network parameters: $W^{new} = W^{current} - \eta \frac{\partial L_{multi}}{\partial W}$;
  end
until $(n = N)$ && $(k = K)$;
Return W;

function. $S_c \in R^M$ and $S_\Gamma \in R^M$ denote the c^{th} and Γ^{th} column of the softmax weight matrix S, respectively. The softmax function is as follows:

$$L_{softmax} = \beta(I) = P(z = c|I) = \frac{exp({S_c}^\top I)}{\sum_\Gamma exp({S_\Gamma}^\top I)} \tag{3.12}$$

In order to jointly train the network with both triplet loss and identification loss, the overall multi-loss training function is as follows:

$$L_{multi} = L_{triplet} + L_{softmax} \tag{3.13}$$

We mention the training approach in Algorithm 1.

3.2 DJI01 Dataset

We introduce DJI01 dataset for video based person re-identification. It includes aerial videos captured through two drones in various challenging situations. It contains 200 identities, with video sequence length varying from 32 to 3000 frames. We show a comparison with the existing datasets in Table 3.1.

Table 3.1: *Comparison of DJI01 parameters with existing datasets (**IDs** denote number of individual identities)*

Datasets	IDs	Type	Cameras
PRID-2011[1]	200	Static	2
iLIDS-VID[3]	300	Static	2
MARS[2]	1261	Static	6
MRP[45]	84	Drone	1
DRHIT01[134]	101	Drone	1
DJI01	200	Drone	2

3.2.1 Data acquisition and processing

Data Collection is done using DJI Phantom 3 and DJI Phantom 4 Quadcopters. The dataset is captured in out-door environment with different backgrounds, altitudes ranging from 3m to 35m, frame rate of 24 fps or 60 fps and resolution is 1280x720. GMMCP tracker [135] is used to segment the person of interest from the video.

3.2.2 Challenges

In our new dataset, we collect rich information with large apperance variation for every single person. The camera's motion and orientation leads to change of view point, which adds challenges during re-identification. Further, the camera motion can also cause blurring effect. Thus, the performance of algorithm may degrade in videos captured using drones. We show different examples of DJI01 in Figure 3.3. It shows challenges of scale variation, pose variation, view variation, different altitude, occlusion, camera motion and illumination variation. In addition, there are 10 same people captured at different days with all the above mentioned challenges. Additionally, 10 subjects illustrate the challenge of people with very similar clothes.

3.2.3 Evaluation Protocol

Several approaches have been used for evaluating re-identification performance. We use Cumulative Match Characteristic (CMC) accuracy to evaluate the performance of person re-identification. All CMC accuracies are given in percentage. The evaluation methodology used here is as follows. The set is split evenly into a training and test set. The CMC for the test set is calculated by selecting a probe video and matching with a gallery video. This provides ranking

Figure 3.3: *Various subjects showing different challenges in DJI01. First row: First subject refers illumination variation, low resolution and clutters; second subject represents scale variations and view variability; third subject shows occlusion and illumination changes; and the fourth one shows illumination changes and camera motion. Second row: First subject shows scale variation, pose variation and blurring due to camera motion; second subject shows occlusion, altitude, scale and pose variation; third subject shows illumination changes, and the fourth one shows pose and view variations. Third row: First two subjects show same persons at different days, and rest two illustrate persons with similar clothes.*

for every video in the gallery w.r.t the probe. This procedure is repeated for every probe video. The CMC is then the expectation of finding the correct match in the top n matches.

3.3 Experiments

We present a comprehensive evaluation of our approach by comparing it to state-of-the-art methods [2, 20, 21, 43, 49, 51, 63, 67, 136, 137]. In [67], the authors use two stream CNN, where each stream is a siamese network to learn spatio-temporal features separately. ASTPN [43] takes the advantage of attention mechanism to extract features from informative frames. In [2], the authors use Alexnet to extract features and metric learning to compute the similarity. [20] learns an intra-video and inter-video distance metric from the training videos. In [21], a recurrent convolutional network (RCN) is used with temporal pooling. STA [51] builds a spatio-temporal appearance representation for person re-identification. Recurrent feature aggregation (RFA) network [63] is based on LSTM. [49] uses 3D HOG, color and LBP features, and learn a distance metric for matching. AFDA [136] uses the representative data samples to learn a feature subspace maximizing the Fisher criterion. [137] learns a single dictionary to represent both gallery and probe images in the training phase. TSCN [68] uses different streams to learn different aspects of feature maps for attentive spatio-temporal fusion of video, and then merges them together to study some union features. In DCF [69] , a Multi-Scale Context-Aware Network (MSCAN) is designed to learn powerful features over full body and body parts, which can well capture the local context knowledge by stacking multi-scale convolutions in each layer. [70] propose SATPP model to leverage the rich visual-temporal cues for

feature learning.

3.3.1 Datasets

A brief description about PRID-2011 [1], iLIDS-VID [3] and MARS [2] datasets is given in Table 3.1. PRID-2011 is captured in uncrowded outdoor environment with stark difference in illumination, background clutter and less occlusions. It contains 200 identities, with video sequence length varying from 5 to 675 frames. iLIDS-VID dataset is more challenging due to occlusions, illumination changes and viewpoint variations. The video sequence length varies from 23 to 192 frames. MARS is much larger dataset when compared to others and sequence length varies from 32 to 20,000.

3.3.2 Experimental set-up

Our architecture is implemented using Torch. All the experiments are performed on Nvidia GTX 1080 GPU. It takes approximately 20 hours for training with 1000 epochs. The values of the hyper-parameters are empirically set. We use the following values, learning rate = 0.001, momentum = 0.9, dropout ratio = 0.6 and the feature embedding-space dimension is 128. For these experiments, each dataset is randomly split into 50% for training and 50% for testing. All experiments are repeated 10 times with different train/test set to ensure stable results. For the MARS dataset, we use the provided fixed training and test set, containing 631 and 630 identities respectively. The source code[1] and dataset[2] are available.

3.3.3 Performance Comparison

PRID, iLIDS, DJI01 and MARS Results

We report the results in terms of CMC-1-5-10-20 accuracies in Table 3.2. On PRID-2011, we achieve 79% CMC-1 accuracy, 96% CMC-5 accuracy, 98% CMC-10 accuracy, and 99% CMC-20 accuracy. We observe that the proposed algorithm gives better results when compared to several popular algorithms in most of the cases.

On iLIDS-VID, CMC-1 accuracy is 65%, CMC-5 accuracy is 87%, CMC-10 accuracy is 93% and CMC-20 accuracy is 98%. Here again, we observe that the performance of the proposed algorithm is significantly better when compared with other schemes. The CMC CMC-1-5-10-20 accuracies are reported in Table 3.3.

The results on MARS dataset are reported in Table 3.4. We observe that the performance of the proposed algorithm significantly outperforms other existing algorithms in both CMC-1 and CMC-10 accuracies.

The results on DJI01 are reported in Table 3.5. A CMC-1 accuracy of 64% is obtained. We also test with other algorithms [2, 21, 43]. We can conclude that our Network outperforms

[1] https://github.com/kajal15003/CARF-Net.git

[2] https://drive.google.com/drive/folders/1HGq5jeIlkJoxYju3zFdzZvKneN$_e pKTK?usp = sharing$

Table 3.2: *Comparison of our proposed approach (CARF-Net) with the state-of-the-art on PRID-2011 [1]*

Methods	CMC-1	CMC-5	CMC-10	CMC-20
DTDL[137]	40.6	69.7	77.8	85.6
TDL[49]	56.74	80	87.64	93.54
RFA[63]	58.2	85.8	93.4	97.9
STA[51]	64	87	90	92
RCN[21]	70	90	95	97
SI^2DL[20]	76.7	95.6	96.7	98.9
ASTPN[43]	77	95	**99**	99
IDE[2]	77.3	93.5	–	**99.3**
TSS-CNN [67]	78	94	97	99
CARF-NET	**79**	**96**	98	99

Table 3.3: *Comparison of our proposed approach (CARF-Net) with the state-of-the-art on iLIDS-VID [3]*

Methods	CMC-1	CMC-5	CMC-10	CMC-20
DTDL[137]	25.9	48.2	57.3	68.9
STA[51]	44	72	84	92
SI^2DL[20]	48.7	81.1	89.2	97.3
RFA[63]	49.3	76.8	85.3	90.0
IDE[2]	53.0	81.4	–	95.1
TDL[49]	56.33	**87.60**	91	96
SATPP [70]	56.67	78.67	90.00	96.67
RCN[21]	58	84	91	96
TSS-CNN [67]	60	86	93	97
ASTPN[43]	62	86	**94**	**98**
CARF-Net	**65**	87	93	**98**

other networks by a significant margin for both CMC-1 and CMC-5 accuracy.

The average CMC-1 accuracy of our algorithm across all four datasets shows a significant improvement of 6.1%, 9% and 14.25% over IDE [2], ASPTN [43] and RCN [21] respectively. In addition, we use XQDA [138] metric learning method for similarity evaluation in feature vectors. The XQDA algorithm learns a discriminant subspace as well as a distance metric simultaneously, and it is able to perform dimension reduction and select the optimal dimensionality. With incorporation of XQDA metric learning there is an improvement of CMC-1 results: 1.20% in PRID-2011 [1], 1.86% on iLIDS-VID [3], 0.90% on MARS [2] and 2.40% on DJI01.

Complexity

We compare the number of parameters used in our network and in other algorithms in Table 3.10. CARF-NET uses only 0.75M parameters. On the other hand, methods like IDE [2], LuNet [37], PBF[139] and TLST[25] uses deep networks such as CaffeNet, ResNetV2, ResNet-50 and 3D-VGGNet require extremely high number of parameters which increases the

Table 3.4: *Results on MARS [2] Dataset. RK:Re-rank*

Methods	CMC-1	CMC-5	CMC-10	CMC-20
RCN[21]	40	64	70	77
ASTPN[43]	44	70	74	81
TSCN[68]	45.6	72.4	75.4	82.6
IDE [2]	65.0	81.1	–	88.9
SATPP [70]	69.69	84.65	89.34	92.77
SFT [65]	70	**90**	–	97
IDE(R)+(RK)[66]	70.51	–	–	–
DCF [69]	71.77	86.57	–	93.08
CARF-Net	**74**	83	**92**	**99**

Table 3.5: *Results on our Drone (DJI01) Dataset*

Methods	CMC-1	CMC-5	CMC-10	CMC-20
RCN[21]	55	62	83	92
IDE[2]	59	65	82	84
ASTPN[43]	63	70	**85**	**98**
CARF-Net	**64**	**74**	83	95

complexity of model. In addition, CARF-Net achieves all results with a feature size of 128, while others like IDE [2] requires 1024 dimensional features, PBF[139] and TLST[25] uses 4096 dimensional feature size. Thus, our model has twofold advantage of being lightweight while learning features with high discriminative ability in a much lesser dimension. We also observe that LuNet achieves an accuracy of 75.56% on MARS, whereas our network attains a comparable accuracy of 74.9% with metric learning. These observations are quite significant as our network is shallow and uses approximately 15% of the parameters used by LuNet. Thus, based upon the above mentioned experimental results, we can conclude that the attentive mechanism can efficiently utilize spatial and temporal human appearance to learn highly discriminative representations.

Table 3.6: *Comparison with deep networks in terms of number of parameters, feature size and CMC-1 accuracy on MARS dataset*

Method	Parameters	Feature Size	CMC-1
IDE (CaffeNet)[2]	60M	1024	70.51
LuNet (ResNetV2)[37]	5M	128	75.56
PBF + ResNet-50[139]	25M	4096	72.64
PBF + VGG16 [139]	134M	4096	67.27
TLST (3D-VGG)[25]	17M	4096	61.66
CARF-Net	0.75M	128	74
CARF-Net+XQDA	0.75M	128	74.90

Cross Dataset Testing

We discuss the comparison results in Table 3.7. Here, we train the system on i-LIDS-VID and test on PRID-2011. It is evident that CARF-Net performs better except for being slightly inferior to TRL method at CMC-5. Thus, we can conclude that our model generalizes well to other datasets.

Table 3.7: *Comparison of cross Dataset Testing results, Test Data- PRID- 2011, Training data - iLIDS-VID*

Methods	CMC-1	CMC-5	CMC-20
ASTPN [43]	30	58	85
RCN[21]	28	57	81
TRL[19]	29.50	**59.40**	82.20
CARF-NET	**36**	55	**87**

Fixed Video Sequence Length

We also analyse the performance of algorithm on iLIDS with fixed number of frame length sequence for both probe and gallery videos. We use a length of 1, 8, 16 and 32 frames. These results are shown in Table 3.8. Here, we observe a CMC-1 accuracy of 20 % for 1/1 sequence length, 38% for 8/8 (probe/gallery length) sequence length, 46% for 16/16 sequence length and 55% for 32/32 sequence length. It also indicates that as we decrease probe and gallery sequence length, re-identification accuracy also reduces as smaller sequences will have less number of discriminative frames and insufficient temporal information.

Table 3.8: *Comparison on iLIDS-VID [3] (CMC-1 Accuracy) for sequence length of 1, 8,16 and 32 frames*

Seq. Length	1/1	8/8	16/16	32/32
ASTPN[43]	16	35	**48**	**59**
RCN [21]	14	28	36	44
CARF-Net	**20**	**38**	46	55

Ground to aerial video person Re-ID

We also study person re-identification in a setting where the data is captured with a moving aerial and a static ground-based camera. We collect the data of 30 persons. We use trained model on DJI01 to test this dataset. The cross dataset testing results are reported in Table 3.9. It is more challenging due to large variations in the characterstics of both cameras like viewing angle, color, resolution etc.

Retrieval Results on DJI01 dataset

In Fig 3.4, we visualize retrieval results on the DJI01 dataset using the existing algorithms [2, 21, 43] and compare it against the proposed algorithm CARF-Net.

Table 3.9: *Results on our Drone Dataset (DJI01) for Ground to aerial video re-identification*

Methods	CMC-1	CMC-5	CMC-10	CMC-20
ASTPN[43]	**58**	69	85	**98**
IDE[2]	51	65	78	94
RCN[21]	45	56	75	90
CARF-Net	56	**72**	**88**	**98**

The videos in the first column are the query videos. The retrieved videos are sorted according to the similarity scores from left to right (from second column till last). Red dashed boundary indicates a negative match and blue shows a positive match. We randomly take probe videos as query and feed as an input to all the above algorithms and retrieve the corresponding gallery videos. In Fig 3.4(a), we show results of IDE [2]. We show the results for RCN [21] in Fig 3.4(b). ASTPN [43] results are given in Fig 3.4(c). We can see that RCN and ASTPN is covering correct results for one of the queries under Top-5. In Fig 3.4(d), we show our results. We observe that in all the scenarios the algorithm covers ground truth in Top-5 retrieval results. In addition, second row shows that the proposed algorithm is robust enough to discriminate people wearing similar clothes.

Ablation Study

We conduct ablation study of some important factors of our method. We compare CMC results of all the variants of CARF-Net. The following are the different variants of our model. C-Net refers to the spatial pooling network, CA-Net refers to attentive spatial pooling network and CAR-Net refers to attentive spatial pooling network with RNN. CARF-Net stands for the combination of CAR-Net and fusion of the spatial and temporal features. The results are shown in Fig 3.5. The average CMC-1 accuracy of CA-Net over all four datasets shows a significant improvement of 9.25% over C-Net, CAR-Net by 14.20% over CA-Net and CARF-Net by 4% over CAR-Net. Our experimental results show that spatio-temporal fusion information obtained using attention mechanism is an important cue for person re-identification and is efficiently captured using the proposed framework (CARF-Net). We believe that the attention mechanism, multiple fusion operations across space and time and carefully chosen loss function exploit the essential regions from the frames which primarily contribute to increase in accuracy.

Robustness to Gaussian Noise

In order to test the robustness against noisy samples, we add Gaussian noise with zero mean and variance between 0.0001 to 0.003. We test noisy examples with the proposed algorithm and existing works [2, 21, 43]. The results are shown in Fig 3.6 over all the four datsets. We can observe that for CMC-1 accuracy, CARF-Net performs better on PRID-2011 and DJI01, whereas ASTPN [43] performs better on iLIDS-VID and IDE [2] performs better on MARS dataset. The average CMC-1 accuracy across all four datasets is 44.00% for CARF-Net, 40.50% for ASTPN [43], 35.25% for IDE [2] and 28.25% for RCN [21].

Since the vulnerability of RGB images against noisy samples is well known, we investigate

(a) Results on DJI01 using IDE [2]

(b) Results on DJI01 using RCN [21]

(c) Results on DJI01 using ASTPN[43]

(d) Results on DJI01 using CARF-Net

Figure 3.4: *Visualization of Top-5 retrieval results on DJI01.*

the performance degradation due to optical flow stream only. We perform experiments with noisy RGB and noisy optical flow as well as with noisy RGB and clean optical flow. Here, noisy optical flow is the optical flow obtained from noisy RGB frames and clean optical flow is obtained from clean RGB frames. In the former case, average accuracy across all four datasets is 44.00% for CARF-Net and 40.50% for ASTPN [43]. In the later case, average accuracy is 46.25% for CARF-Net and 41.25% for ASTPN [43]. We observe that the accuracy change between the two cases is low. This suggests that optical flow inherently provides robustness when noise is added spatially. Optical flow gives temporal information and may not be as vulnerable as noisy RGB frames. Further, we also see that the accuracy increases with clean optical flow is highest for the case of CARF-Net. This may also indicate the fact that two stream networks provide an advantage over single stream networks.

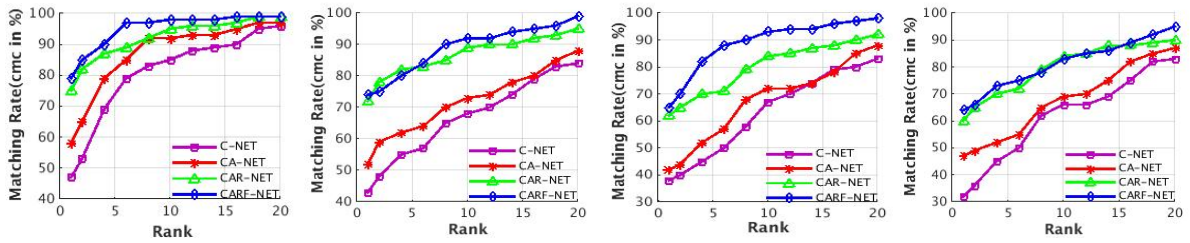


Figure 3.5: *Ablation study results on PRID-2011 [1], MARS[2], iLIDS-VID [3] and DJI01 (from left to right).*

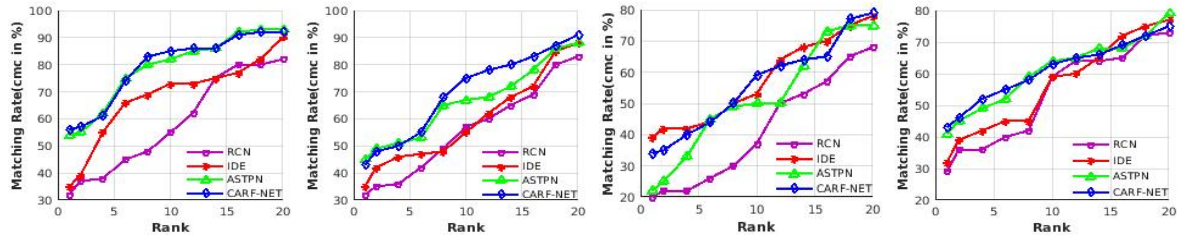


Figure 3.6: *Noise Robustness results on PRID-2011 [1], iLIDS-VID [3], MARS[2], and DJI01 (from left to right).*

T-SNE Plots

In order to qualitatively understand each component of CARF-Net, we analyse learned embedding from C-Net, CA-Net, CAR-Net and CARF-Net using T-SNE [140], a dimensionality reduction technique to visualise high-dimensional embeddings. We obtain the two resulting dimensions which we can visualise by creating a scatter plot of the two dimensions and coloring each sample by its respective label. We plot 220 samples for 20 identities from MARS [2] dataset. We extract four types of embeddings from C-Net, CA-Net, CAR-Net and CARF-Net in Fig 6.4 (from left to right). **C-Net:** we can see that most of the dissimilar points cluster together incorrectly. **CA-Net:** we can see that similar points gets more closer as compared to C-Net. It also shows the effect of SPP layer. **CAR-Net:** we can easily differentiate the clusters as they are well aligned and well separated, though 40% clusters still contains samples of different identities. **CARF-Net:** we can see that the samples are very clearly clustered in their own little group and more than 90% of similar points are correctly clustered.

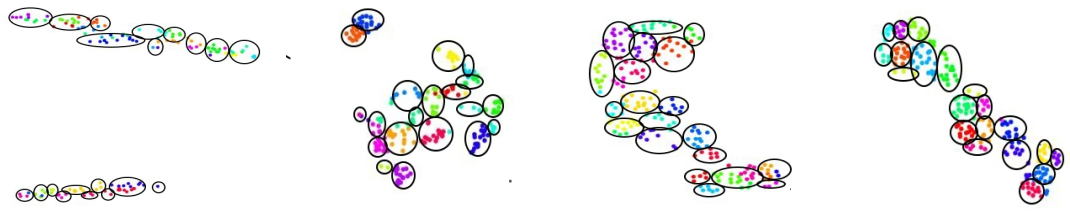


Figure 3.7: *Visualizations of the MARS[2] Embeddings by t-sne: Embedding from C-Net, CA-Net, CAR-Net and from CARF-Net (from left to right).*

3.4 Discussion

In one of our works, we have employed Conv3D to extract spatio-temporal features to accomplish video based person Re-ID. Here, we propose a spatio-temporal transfer learning (TLST) approach using 3D-CNN where we address the issue of insufficient labelled data and transfer the knowledge. We use a pre-trained 3D-CNN model of Sports-1M [141] dataset and perform fine-tuning on multiple domain datasets such as PRID-2011, iLIDS-VID, MARS and an aerial video dataset simultaneously. Learning features from multiple domain data is of significant value because of large variation which otherwise is not possible to obtain from small individual datasets. In our experiments, we show that the fine-tuned transferred features encode robust representations and enhance the re-identification accuracy. In addition, we analyse the network's robustness against adversarial examples and show that the proposed 3D-CNN network has better resilience compared to 2D-CNN used in most of the existing algorithms.

Our 3D-CNN architecture is inspired from 3D-VGGNets [60]. The architecture has 8 convolutional layers, 5 max-pooling layers, and 3 fully connected layers, followed by softmax-loss layer. In [60], authors point out that a homogeneous architecture with 3x3x3 convolution kernels in all layers perform best for 3D-CNN. In our 3D-CNN, we also apply $3\times3\times3$ convolutional Kernels and $2\times2\times2$ pooling kernels. All 3D convolutional kernels are $3\times3\times3$ with stride 1 in both spatial and temporal dimensions. All pooling kernels are $2\times2\times2$, except for 1st pooling layer kernel which is $1\times2\times2$. We use Rectified linear unit (ReLU) as an activation function to accelerate the convergence of stochastic gradient descent. The convolutional layers are followed by three fully connected layers. First two fully connected layers have 4096 output units. We use the second fully connected layer for feature extraction. It gives feature vectors corresponding to the appearance and motion of a person in the video.

We compare the number of parameters used in CARF-Net and 3D-CNN network in Table 3.10. CARF-NET uses only 0.75M parameters. On the other hand, TLST[25] uses deep networks, 3D-VGGNet which requires extremely high number of parameters which increases the complexity of model. In addition, CARF-Net achieves all results with a feature size of 128, TLST[25] uses 4096 dimensional feature size. Thus, the CARF-Net model has twofold advantage of being lightweight while learning features with high discriminative ability in a much lesser dimension. Thus, based upon the above mentioned experimental results, we can conclude that the attentive mechanism can efficiently utilize spatial and temporal human appearance to learn highly discriminative representations.

Table 3.10: *Comparison with deep networks in terms of number of parameters, feature size and CMC-1 accuracy on MARS dataset*

Method	Parameters	Feature Size	CMC-1
TLST (3D-VGG)[25]	17M	4096	61.66
CARF-Net	0.75M	128	74

This chapter covered novel method for video-video matching along with the details of novel air borne video dataset under homogeneous modality. The main focus of the algorithm was on encoding the robust spatio-temporal fusion features via light-weight shallow network. In the next chapter, we will discuss the proposed method for RGB-RGB Re-ID.

Chapter 4

RGB-RGB based Person Re-ID

In this chapter, we describe our proposed model for homogeneous image based person Re-ID, named as HDRNet in section 4.1. Our model comprises of an encoder-decoder architecture, where the primary goal is to learn efficient embeddings for person Re-ID. We use multiple decoders to reconstruct images from learned feature representations to further enhance the robustness of feature embeddings. We further describe our batch median based sampling strategy in section 4.1.2 followed by test set augmentation in section 4.1.3. We also report qualitative and quantitative results of comparison with other popular Re-ID algorithms.

4.1 Network Architecture

The overall network architecture of the HDRNet is illustrated in Figure 4.1. The encoder architecture is based on that of ResNet-50 [142] which has been widely used as the base network in many vision applications. We discard the last layer and add two fully connected layers denoted by FC-1 and FC-2. The last fully connected (FC-2) layer has 128 units. This 128-D output is taken as the feature representation for describing the visual appearance of the person. This layer is also used as input to three parallel decoders which reconstruct output at three different resolutions (128×256, 64×128, and 64×32). The goal of augmenting the network with decoder is three-fold. First, it has been shown that additional tasks improve the accuracy of the primary task [143]. Thus, the decoder helps the encoder to learn a discriminative feature representation that generalises better. Second, in case of multiple decoders, if the encoded features can reconstruct the input at different resolutions, the features shall be invariant to image resolution. Third, it helps in circumventing the challenges of pose, occlusion and illuminance by reconstructing the samples with uniform pose and illuminance while removing occlusion. Thus, the gradients that flow during backpropagation, enhance the feature robustness against these challenges. The decoder network is a shallow network which consists of 3 convolution layers, named as CONV-1, CONV-2, and CONV-3 and upsampling layers denoted by U at different resolutions. The output of FC-2 layer is reshaped, upsampled and given as input to CONV-1. Similarly, the output of CONV-1 is upsampled and given as input to CONV-2. After performing last upsampling, the reconstructed output is obtained from CONV-3.

The FC-2 layer output is also used to compute the $N \times N$ Euclidean distance matrix using the N samples. Then, we apply hybrid sampling strategy to mine triplets. We discuss this strategy under subsection 4.1.2. We train the network using triplet verification and mean squared error loss.

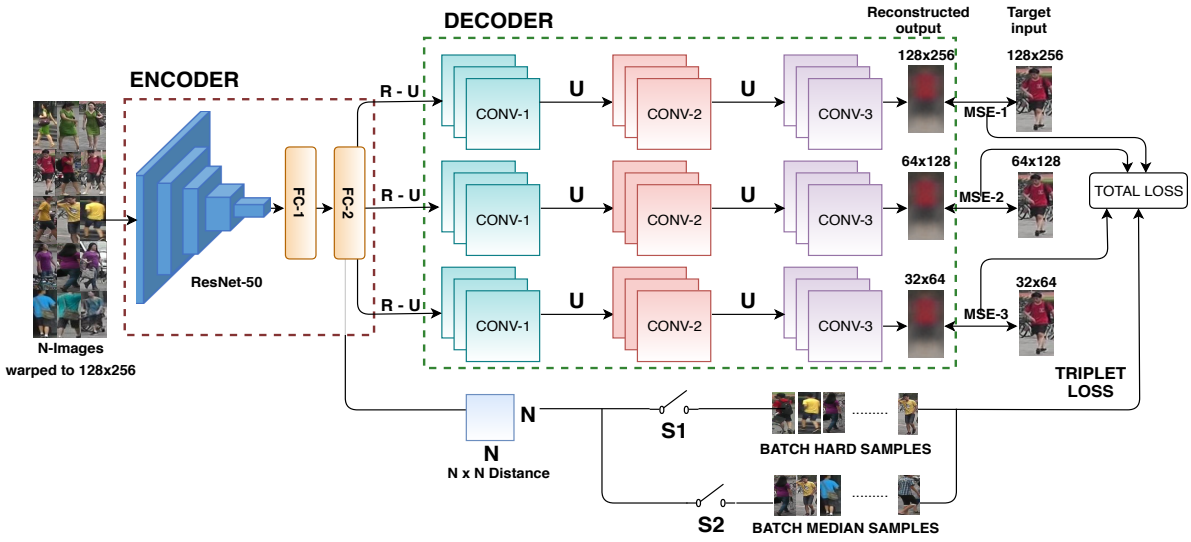


Figure 4.1: *Proposed framework HDRNet. The input to the ResNet-50 network comprises of* ***N*** *color images of size* 128×256. ***FC-1*** *and* ***FC-2*** *denotes two fully connected layers. Output of the last fully connected layer is given as an input to the multi-resolution decoder.* ***R-U*** *refers to the reshape operation followed by upsampling.* ***CONV-1***, ***CONV-2*** *and* ***CONV-3*** *are convolutional layers.* ***U*** *denotes the upsampling operation.* ***MSE-1***, ***MSE-2*** *and* ***MSE-3*** *refers to mean squared error at three resolutions. Distance metrics is represented by* ***N X N*** *Distacne between* ***N*** *images.* ***S1*** *and* ***S2*** *represent switches to enable batch hard and batch median sampling strategy, respectively.* ***S1*** *is ON for first 1K iterations only, whereas,* ***S2*** *is ON for the remaining iterations during training. Total Loss is addition of Triplet and MSE losses.*

4.1.1 Model Learning

Input

We randomly sample P classes (person identities) and sample K images of each class (person), thus resulting in a batch of PK input images.

Learning Objectives

Given an input image x_i, the encoder outputs a feature representation $O_E(x_i)$ where, $O_E(.)$ denotes the output of the encoder. The triplet verification loss function is as follows:

$$L_T = \frac{1}{N}\sum_{i=1}^{N}\Big[\|O_E(x_i) - O_E(x_i^+)\|_2^2 - \|O_E(x_i) - O_E(x_i^-)\|_2^2 + \alpha\Big]_+ \tag{4.1}$$

where α is a positive constant, x_i is the anchor, x_i^+ is the positive sample, x_i^- is the negative sample, N is the number of triplets and $[.]_+$ is the hinge function.

Further, encoder is followed by multiple resolution decoder which reconstructs the input at different resolutions. In our case, the target output images are $\hat{x_i}^s$, where $\hat{x_i}^1$ is the original resolution input image, $\hat{x_i}^2$ is one-fourth resolution of input image, and $\hat{x_i}^3$ is one-sixteenth resolution of input image. Hence, the loss of the decoder is given by L_D as follows:

$$L_D = \sum_{s=1}^{S} \|\hat{x_i}^s - O_D^s(O_E(x_i))\|_2^2 \tag{4.2}$$

where S is the number of scales whose value is taken as 3. $O_D^s(.)$ is the output of the decoder.

We aim to obtain generalizable and discriminative features for Re-ID, thus, our focus is not on high quality reconstruction of images. We train the model with weighted loss function given by, $L_T + \lambda L_D$. *Further, the weighting factor* λ *is experimentally tuned to* 10^{-6}.

4.1.2 Hybrid Sampling Strategy

In this section, we discuss a novel sampling strategy named as Hybrid sampling strategy, which is a sequential application of batch hard strategy and batch median strategy. The batch hard [37] samples the positive example using a max operator (100^{th} percentile) and negative example using a min operator (0^{th} percentile). However, it has been shown that median operator (50^{th} percentile) is robust to outliers while max and min operators are adversely affected by outliers [144]. Inspired by this observation, we *propose batch median sampling, where we select the median positive and the median negative samples within the batch while forming the triplets.* First, we use batch hard sampling for 1K iterations and then apply batch median sampling for the 24K iterations. We qualitatively analyze the effect of batch hard as well as hybrid sampling strategy in Figure 4.3. In Figure 2(b), we find that the maximum distance for positive samples after applying hybrid strategy for training is about 31% lower compared to Figure 2(a) where only batch hard strategy is used. Further, the number of samples achieving this is similar in both the cases. *Thus, highlighting the fact that similar number of positive samples in our strategy achieve much lower distance.* Similarly, in case of negative samples, we find that our hybrid strategy achieves a maximum distance of 17% higher compared to batch hard strategy. [1]

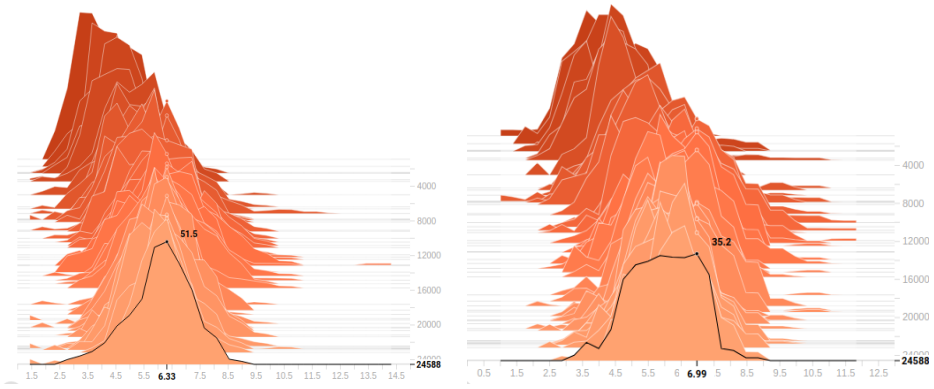
4.1.3 Test Set augmentation

We propose a novel test set augmentation technique with the reconstructed images in single and multiple query test setting. In case of Market-1501 and DukeMTMC-reID, we augment the test image using reconstructed image $O_D^1(O_E(x_i)$. We compute the average of feature embeddings of both images and then determine the cumulative matching characteristic (CMC) accuracy.

4.2 Discussion

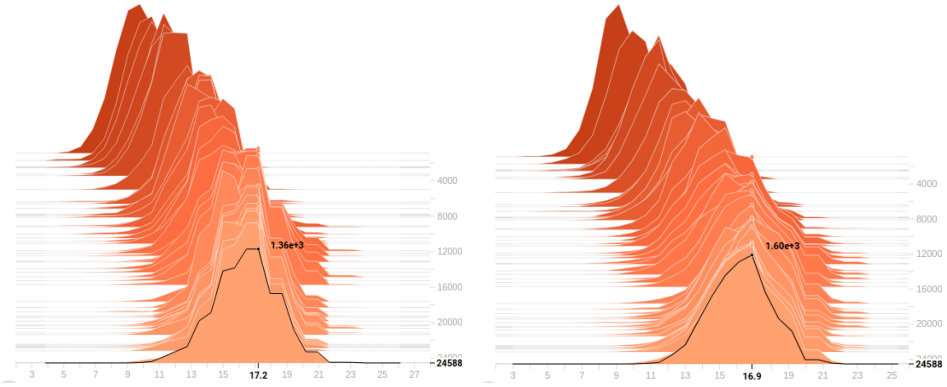
In this section, we discuss the network' robustness against specific challenges. *As our model is trained with triplets of the same identity having different view angles, poses, lighting conditions,*

[1] These plots are created during training using tensor-board. Please zoom-in for better clarity.



(a) (b)

Figure 4.2: *Training plots for Market-1501: Positive Sample Distance Embeddings (a) Batch hard (b) Hybrid strategy. Please refer to the web version for clarity.*



(a) (b)

Figure 4.3: *Training plots for Market-1501: Negative Sample Distance Embeddings (a) Batch hard (b) Hybrid strategy. Please refer to the web version for clarity.*

and occlusions, we expect that the features are mapped to a common subspace and reconstruction using these features would have all the distractions removed while preserving the identity relevant information. Figures 4.4 and 4.5 show the reconstructed images corresponding to the input images of test set.

4.2.1 Pose Invariance

In the following, we provide reconstruction examples and demonstrate that the network learns embeddings invariant to pose. In Figure 4.4 (first row), when person is moving from one camera to another, pose gets changed due to orientation of camera. However, reconstructed examples maps all the different poses to same frontal pose and removes pose challenge completely. We can observe a similar effect in Figure 4.4 (second row). It is also evident that some important visual attributes are preserved, for example, the color of the T-shirt.

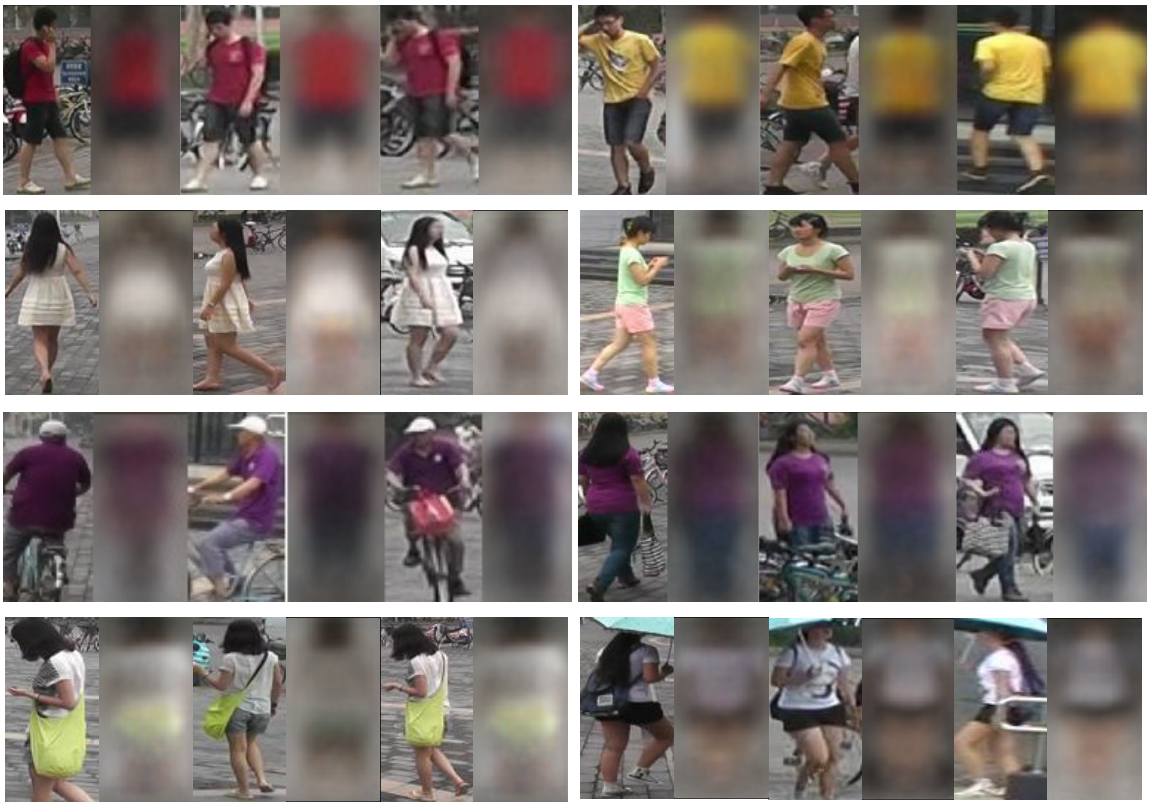


Figure 4.4: *Original and reconstructed images: For both rows, first, third and fifth column shows original samples with different poses and occlusion. Second, fourth and sixth column shows the reconstructed images. Please refer to the web version for clarity.*

4.2.2 Occlusion Invariance

In Figure 4.4, we observe that the background clutter as well as occlusion is removed. For example, the bicycle, handbag, umbrella and backpack are eliminated (in third and fourth row). Thus, we observe that these features are only dependent on the person to be re-identified.

4.2.3 Illumination Invariance

In Figure 4.5, we can see how colour changes when we move from one camera to another (first column and third column represents input images from two different cameras with illumination variation). In the first row, the person' T-shirt color changes from blue to green which completely alters the appearance and poses a big challenge in Re-ID. However, from the reconstructed images we can visualize that different colours map to the same colour by keeping identity preserving information. However, the proposed network is not completely addressing the challenge of illumination invariance as this is a huge challenge. The illumination variation challenge can be further explored as a future aspect.

Thus, generated images suggest that having the generative decoder would encourage the learned feature to focus on the identity-related appearance rather than environment-related

Figure 4.5: *First and third column: Samples from different cameras describing major illumination challenge. second and fourth column : reconstruction results showing robustness against illumination change. Please refer to the web version for clarity.*

distractions such as background, occlusion, and pose variation, thus helping the discriminative Re-ID task.

4.3 Experiments

In this section, we describe datasets used and their evaluation protocols. Further, we discuss ablation study to evaluate the effectiveness of each component in our proposed model. We also compare results of HDRNet with state-of-the-art methods. We also perform experiments to investigate the robustness against individual challenges.

4.3.1 Datasets and Evaluation Protocol

We evaluate our model on the following three datasets: Market-1501 [8], DukeMTMC-reID [9], and CUHK03 [7]. We follow the official protocols to evaluate the proposed HDRNet model. On Market-1501, we follow [8] by splitting all 1501 identities into two halves of 751/750 for training and testing, respectively. Further, we evaluate both single-query and multi-query test settings. On DukeMTMC-reID, we use the standard training/test split (702/702) [9] and evaluate single-query test setting. On CUHK03, we consider two different settings. In the first evaluation, we use 1367/100 train/test setting [7]. The experiment is repeated 20 times with random splits and the average accuracy is reported. In the second evaluation, we use 767/700 train/test setting [66]. The single-shot evaluation setting is applied for both cases. The cumulative matching characteristic (CMC) accuracy is computed to evaluate all the methods. In addition, we calculate the mean average precision (mAP).

4.3.2 Implementation Details

The model is implemented on TensorFlow with NVIDIA GTX-1080 GPU. We use the ResNet-50 architecture as encoder and weights provided by He et al. [142]. All the images are scaled to 128×256 pixels. Data augmentation includes random horizontal flipping and cropping. The 0.5 margin is used for triplet loss, number of iterations is set to 25K, and the mini-batch size is 22, in which each identity has 4 images. The number of parameters and FLOPs used in the network are 26M and 3859.05M, respectively. We use Adam optimizer with default hyperparameter values, an initial learning rate of 3×10^{-4}, and shrink this learning rate by following exponentially decaying training schedule [37] until the convergence is achieved. The source code is available at Github.

4.3.3 Ablation Study

We compare different combinations of the sub-networks in Table 6.1. We first experiment with the pre-trained networks like AlexNet [145], ResNet-50 [142], Inception [146], and Inception-ResNet [147] as base models and train them on Re-ID datasets. We observe that ResNet-50 gives the best performance. Hence, we keep ResNet-50 as base network. The average CMC-1 accuracy on all datasets is 65.43%. We then augment the base network with multiple decoders. When we add a single decoder to reconstruct the input image with same resolution (128×256), we see an average accuracy improvement of 1.68% over all the datasets. Similarly, we see an improvement of 3.52% and 10.08% when we add decoders for reconstructing the images with quarter as well as one-sixteenth resolution, respectively. Further, we observe that adding more decoders adversely impacts the Re-ID accuracy. In this case, the reconstructed images get sharper and the network may deviate from primary goal of Re-ID. Thus, we find that ResNet-50 with 3 decoders delivers best performance. We name it as DRNet. Then, we use hybrid sampling strategy for ResNet-50 with 3 Decoders and term it as HDRNet. HDRNet gives a rise of 0.44% in single query of Market-1501, 1.61% in detected set of CUHK03 for 767/700 split, and 1.89% in case of DukeMTMC-reID for CMC-1.

Table 4.1: *Ablation Study results on each dataset. R-ResNet-50, D-Decoder, BH-Batch-Hard, HD-Hybrid*

Base Model	Market-1501		CUHK03 (767/700)		DukeMTMC-reID	
	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1
R	59.72	81.25	35.34	45.48	60.87	69.78
R + 1 D	63.85	82.70	37.73	46.32	63.56	72.87
R + 2 D	67.45	83.50	40.13	48.40	68.56	75.67
R + 3 D + BH (DRNet)	72.89	87.38	52.89	57.84	71.32	81.23
R + 3 D + HD (HDRNet)	**73.56**	**87.82**	**54.50**	**58.42**	**74.11**	**83.12**

Table 4.2: *Evaluation on Market-1501 and DukeMTMC-reID. RK: Re-ranking, TA:Test Augmentation, '-': Results are unavailable.*

Methods		Market-1501				DukeMTMC-reID	
		Single Query		Multi Query		Single Query	
		mAP	CMC-1	mAP	CMC-1	mAP	CMC-1
Attention	HA-CNN [91]	75.70	91.20	82.80	93.80	63.80	80.50
	Mancs [82]	82.30	93.10	87.50	95.40	71.80	84.90
	CA^3Net [90]	80.00	93.20	–	–	70.20	84.60
Part-based	GLAD [55]	73.90	89.90	–	–	–	–
	PN-GAN [94]	72.58	89.43	80.19	92.93	53.20	73.58
	Local CNN [148]	77.73	91.51	–	–	66.04	82.23
	AlignedReID [22]	–	91.80	–	–	–	–
	PCB + RPP [88]	81.60	93.80	–	–	69.20	83.30
Multi-Resolution	MS-TriNet [99]	–	45.10	–	55.40	–	–
	MSCAN [69]	57.53	80.31	66.70	86.79	–	–
	DPFL [101]	72.60	88.60	80.70	92.30	60.60	79.20
Global Feature	SVDnet [149]	62.10	82.30	–	–	56.80	76.70
	HAP2SP [150]	69.43	84.59	76.75	90.20	60.64	75.94
	TriNet + RK [37]	81.07	86.67	87.18	91.75	–	–
	AO + RK [81]	83.30	88.66	88.60	92.51	78.19	84.11
	DarkRank [96]	74.30	89.80	81.20	93.70	–	–
	SGGNN [97]	82.80	92.30	–	–	68.20	81.10
Multi-Resolution & Global Feature	**DRNet**	72.89	87.38	79.67	91.32	71.32	81.23
	DRNet+TA	75.38	89.45	83.76	92.01	74.12	82.67
	DRNet+TA+RK	87.15	94.32	91.35	95.25	82.12	86.85
	HDRNet	73.56	87.82	80.61	92.14	74.11	83.12
	HDRNet+TA	77.18	90.65	84.22	93.21	75.34	84.34
	HDRNet+TA+RK	**88.01**	**95.14**	**92.59**	**96.49**	**84.56**	**88.13**

4.3.4 Comparison with State-of-the-Art

Results on Market-1501

We report the CMC accuracies corresponding to DRNet and HDRNet in Table 4.2. In case of HDRNet, the CMC-1 accuracy for single query is 87.82%, for HDRNet + Test-Augmentation (TA), it is 90.65% and for HDRNet + TA + Re-ranking (RK), it is 95.14%. With test augmentation there is a rise of 2.83% accuracy. Hybrid sampling strategy gives a boost of 0.82% compared to batch hard strategy used in DRNet for single query. We can observe a similar trend in case of multi-query setting. While comparing our best model (HDRNet + TA + RK) against the other algorithms which use multiple resolutions such as MS-TriNet and DPFL, we observe an increase of 50.04% and 6.54% respectively for single query setting. In addition, we can see that HDRNet performs very well compared to part based methods including PCB [88] which requires uniform partitioning, GLAD [55] which requires auxiliary part labeling to align parts and AlignedReID [22] which needs local features in addition to global features. Further, HDRNet achieves all results with a feature size of 128, while PCB [88] uses a feature size of 12,288, AlignedReID [22] uses 2048 global feature size, SphereReID [80] uses 1024

Table 4.3: *CMC-1 accuracy on CUHK03. D-detected, L-labelled*

1367/100 split		D	L
Attention	CAN [83]	69.20	77.60
	HydraPlus [89]	–	91.80
Part-based	GLAD [55]	82.20	85.00
	Spindle-Net [87]	–	88.50
Multi-resolution	MSCAN [69]	67.99	74.21
	MuDeep [100]	75.64	76.87
	DPFL [101]	82.00	86.70
Global-feature	TriNet [37]	87.58	89.63
	HAP2SP [150]	88.90	90.40
Multi-Resolution & Global Feature	**DRNet**	88.45	91.34
	DRNet+RK	**91.51**	**94.34**
767/700 split		D	L
Attention	HA-CNN [91]	41.70	44.40
Part-based	Local CNN [94]	51.55	53.83
	PCB +RPP [88]	63.70	–
Multi-resolution	DPFL [101]	40.70	43.00
Global-feature	PAN [151]	36.30	36.90
	SVDnet [149]	40.90	41.50
	AO [81]	54.56	–
Multi-Resolution & Global Feature	**DRNet**	57.84	59.78
	DRNet+RK	63.67	65.17
	HDRNet	58.42	60.23
	HDRNet+RK	**64.12**	**65.78**

feature size. Thus, our model learns features with high discriminative ability but in a much lesser dimension relative to these algorithms.

Results on DukeMTMC-reID

We report the CMC accuracies in Table 4.2. The best competitors are Mancs [82], CA^3Net[90] and AO + RK [81]. Mancs and CA^3Net requires attention model, while AO + RK exploits adversarial occluded samples to train the model. Our best model gives a rise of 3.23%, 3.53 and 4.02% in CMC-1 accuracy as compared to Mancs, CA^3Net and AO + RK.

Results on CUHK03

We report the CMC accuracies for 1367/100 and 767/700 splits in Table 4.3. Among both splits, 767/700 is more challenging and practical. While comparing DRNet + RK against attention models such as HydraPlus [89], we observe an increase of 2.54% for 1367/100 labelled setting.[2] Similarly, our method performs favorably compared to other algorithms in 767/700 setting.

[2]Since 767/700 is more challenging, we prefer hybrid sampling only for this setting.

4.3.5 Robustness Evaluation

Robustness to Pose Variance

In order to investigate the robustness against pose variation only, we manually select images for 350 different persons from Market-1501, CUHK03 and DukeMTMC-reID where the major challenge is pose variation, and other challenges are either absent or mildly present. We report these results in Table 4.4. In case of Market-1501, we can observe a significant rise of about 9.94% for DRNet over ResNet. Thus, it is evident that the multi-resolution decoder significantly boosts the performance. Further, the hybrid sampling strategy leads to an increase of 2.2% over DRNet. Similarly, our method outperforms on other two datasets. This also supports the observation that the sampling strategy effectively supplements the decoder in overcoming the pose challenge.

Table 4.4: *CMC-1 results for pose specific Samples. A : Single-Query for Market-1501, B : CUHK03-Detected (767/700) and C: DukeMTMC-reID*

Methods	A	B	C
ResNet-50 [142]	76.66	53.18	61.27
TriNet [37]	78.37	56.11	65.77
PCB [88]	82.38	68.77	73.23
DRNet	86.60	71.42	82.80
DRNet+RK	88.55	74.01	83.92
HDRNet	88.80	72.71	84.07
HDRNet+RK	**90.02**	**75.04**	**85.92**

Robustness to Partial Occlusion

Partially occluded samples can be defined as artificially generated occluded samples by blacking-out a region in the given image [81]. In the experiments, we chose random regions of about 10% in area of the overall image. We report the CMC-1 results in Table 4.5. The average CMC-1 accuracy on all datasets for HDRNet is 70.58% whereas TriNet gives 53.27%, PCB shows 54.31% and SVDnet obtains 43.64%. This demonstrates that our method is highly robust against occluded samples.

Table 4.5: *CMC-1 results for occluded samples A : Single-Query for Market-1501, B : CUHK03-Detected (767/700) and C: DukeMTMC-reID*

Methods	A	B	C
SVDnet [149]	57.70	23.68	52.35
TriNet [37]	59.70	43.43	54.87
PCB [88]	62.78	42.67	56.23
DRNet	80.59	53.81	74.21
DRNet+RK	82.98	56.78	76.92
HDRNet	80.78	54.87	76.11
HDRNet+RK	**83.34**	**57.12**	**77.88**

Robustness to Low Resolution

Low resolution samples pose a severe challenge as the network trained on moderate resolution inputs may perform very poor when tested with low resolution examples. In practice, these samples can easily occur when a person is captured while being far away from the camera. Therefore, it is necessary to design the network such that it is robust to the challenge of low resolution. We report the accuracy on low resolution images in Table 4.6. Here, we find that in case of single query for Market-1501, CMC-1 accuracy of ResNet-50 gets drastically dropped from 81.25% to 5.86%, whereas, HDRNet performs better by a margin of 41.53%. Similarly, in case of CUHK03 (detected) and DukeMTMC-reID, our method outperforms by a clear margin of 20.24% and 32.79% in CMC-1 accuracy, respectively. Thus, we observe that augmentation of the proposed multi-resolution decoder network brings robustness against the low resolution images.

Table 4.6: *CMC-1 results for low resolution samples of* 64×128*, A : Single-Query for Market-1501, B : CUHK03-Detected (767/700) and C: DukeMTMC-reID*

Methods	A	B	C
ResNet-50 [142]	5.86	15.66	17.02
DRNet	44.32	32.12	47.89
DRNet + RK	46.11	34.88	48.09
HDRNet	45.55	32.77	48.12
HDRNet + RK	**46.89**	**35.90**	**49.81**

We proposed a novel method for RGB-RGB Re-ID in this chapter. The major goal of the algorithm was to learn generalizable features to address the conventional Re-ID challenges. In the next chapter, we discuss the image based heterogeneous Re-ID models.

Chapter 5

RGB-IR based Person Re-ID

This chapter discusses the Spectrum Disentangling Network (SDL) for image based heterogeneous Re-ID model. To better understand the goal of our method, we visually illustrate it in Fig 5.1. Existing RGB-IR models try to learn shared representations ($\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$) by projecting the corresponding RGB-IR pair to a shared subspace. In this case, the spectrum information is not explicitly disentangled from the learned representations. Since it is a challenging task to fill the large domain gap between RGB and IR data, such representations may not be efficient. In order to solve this problem, our network learns the spectrum disentangled features $\boldsymbol{u}$ by effectively removing the spectrum related information ($\boldsymbol{v}_{\mathcal{RGB}}$ and $\boldsymbol{v}_{\mathcal{IR}}$) and retains only identity specific information.

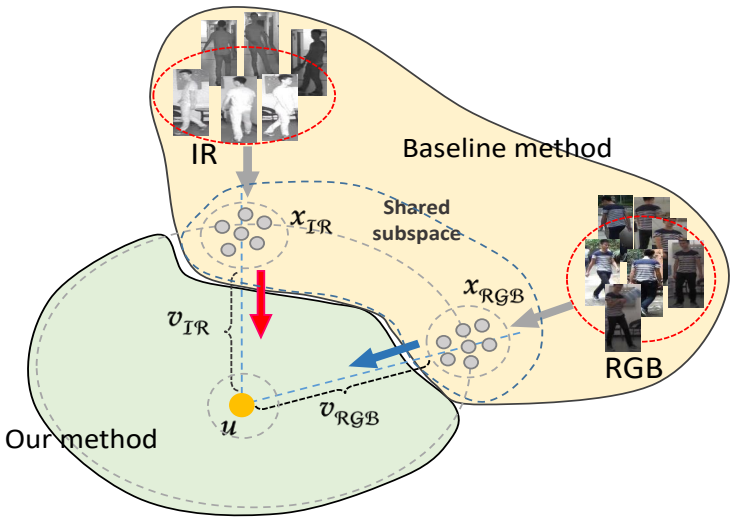


Figure 5.1: Problem formulation and the goal of Spectrum Disentangled representation learning (SDL). *$\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ represents shared features, $\boldsymbol{v}_{\mathcal{RGB}}$ denotes the RGB spectrum features, $\boldsymbol{v}_{\mathcal{IR}}$ represents the IR spectrum features and $\boldsymbol{u}$ denotes the spectrum-disentangled representation. The goal is to remove spectrum related information $\boldsymbol{v}_{\mathcal{RGB}}$ and $\boldsymbol{v}_{\mathcal{IR}}$, respectively from $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ to learn $\boldsymbol{u}$.*

5.1 SDL Network

Our model comprises of a dual-path network as the baseline network to learn shared features between cross modal inputs. The network has two branches – spectrum dispelling and spectrum

distilling. The goal of the spectrum dispelling branch is to keep useful cues of person which are invariant to spectrum. Whereas, the spectrum distilling branch removes the spectrum related information. We apply identity-dispeller over this branch to fool the identity classifiers and extract all the spectrum related information. Further, to disentangle spectrum from the identity discriminative features, we apply a disentanglement loss. We discuss the proposed approach in detail in the following subsections.

5.1.1 Baseline Network

We choose [4] as our baseline. It comprises of two main components: dual-path network for feature extraction (one path for RGB images and the other for IR images) and bi-directional dual-constrained top-ranking loss for feature learning. There are two prominent reasons behind choosing this baseline. First, [4] achieves a significant performance via end-to-end training. Second, the dual path network naturally provides a framework to segregate the spectrum information from identity specific information. The loss for the baseline network is taken from [4]. It is a weighted summation of bi-directional cross-modality, intra-modality top-ranking constraints and identity loss. The baseline loss is denoted as $\mathcal{L}_{\mathcal{B}}$ given in Eq. (5.1).

$$\mathcal{L}_{\mathcal{B}} = \mathcal{L}_{cross} + \mathcal{L}_{intra} + \mathcal{L}_{id}\,. \tag{5.1}$$

where $\mathcal{L}_{cross}$, $\mathcal{L}_{intra}$ and $\mathcal{L}_{id}$ are respectively

$$\begin{aligned}\mathcal{L}_{cross} = &\sum_{\forall_{y=y_j}} max[\rho_1 + D(\boldsymbol{x}_{\mathcal{RGB}_i}, \boldsymbol{x}_{\mathcal{IR}_j}) - \min_{\forall_{y_i \neq y_k}} D(\boldsymbol{x}_{\mathcal{RGB}_i}, \boldsymbol{x}_{\mathcal{IR}_k}), 0] \\ &+ \sum_{\forall_{y_i=y_j}} max[\rho_1 + D(\boldsymbol{x}_{\mathcal{IR}_i}, \boldsymbol{x}_{\mathcal{RGB}_j}) - \min_{\forall_{y_i \neq y_k}} D(\boldsymbol{x}_{\mathcal{IR}_i}, \boldsymbol{x}_{\mathcal{RGB}_k}), 0]\end{aligned} \tag{5.2}$$

Here, $\boldsymbol{x}_{\mathcal{RGB}}$ is an anchor visible image with label y, $\boldsymbol{x}_{\mathcal{IR}}$ is the infrared image and $D(.)$ is the Euclidean distance. i and j subscripts represent the same identity, whereas i and k are different identities. ρ_1 is pre-defined margin. For an anchor visible image $\boldsymbol{x}_{\mathcal{RGB}_i}$ with its label denoted by y_i, we want the distance of its positive IR image $\boldsymbol{x}_{\mathcal{IR}_j}$ should be smaller than the distance between $\boldsymbol{x}_{\mathcal{RGB}_i}$ and the negative IR image $\boldsymbol{x}_{\mathcal{IR}_k}$.

$$\begin{aligned}\mathcal{L}_{intra} = &\sum max[\rho_2 - D(\boldsymbol{x}_{\mathcal{IR}_j}, \boldsymbol{x}_{\mathcal{IR}_k}), 0] \\ &+ \sum max[\rho_2 - D(\boldsymbol{x}_{\mathcal{RGB}_j}, \boldsymbol{x}_{\mathcal{RGB}_k}), 0]\end{aligned} \tag{5.3}$$

where ρ_2 is pre-defined margin. Here, we want to ensure that the hardest cross-modal negative sample should also be far from its corresponding cross modal positive samples.

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{j=1}^{N} \log y_j\,, \tag{5.4}$$

where N is the number of images and y is the predicted probability.

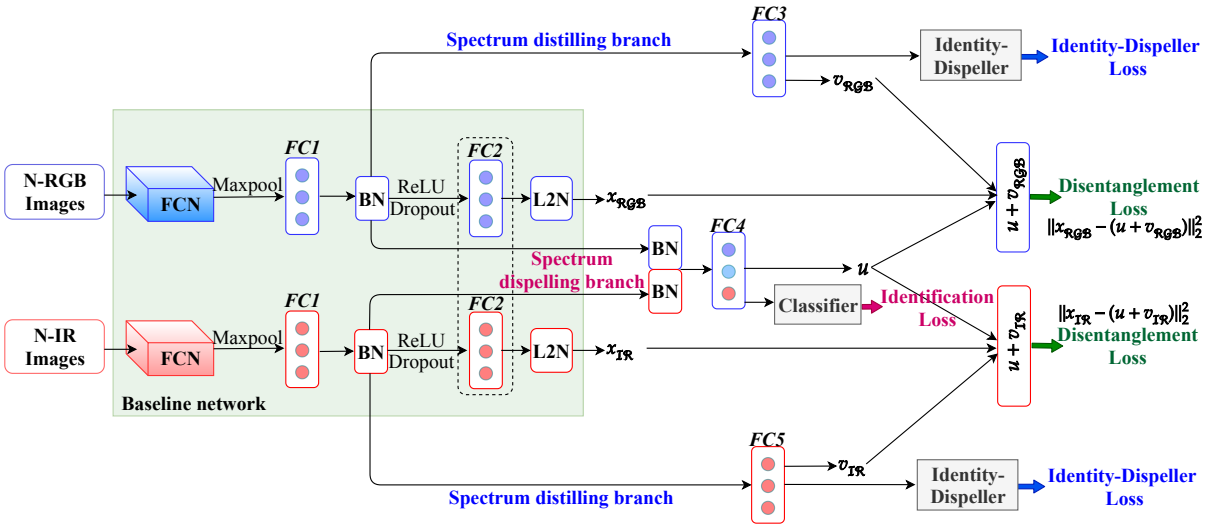


Figure 5.2: The architecture of the proposed SDL network. *The entire network consists of the baseline network [4], extended fully-connected layers and three new kinds of losses. The outputs of FC4 layer is taken as the spectrum-disentangled feature $\boldsymbol{u}$. The outputs of FC3 and FC5 layer are respectively taken as the RGB spectrum feature $\boldsymbol{v}_{\mathcal{RGB}}$ and the IR spectrum feature $\boldsymbol{v}_{\mathcal{IR}}$. The identification loss is used for the feature $\boldsymbol{u}$. The identity-dispeller loss is designed to fool the identity classifier so that it primarily learns spectrum related information. The disentanglement loss distills identity features and removes spectrum features. Note that the module 'BN' stands for batch normalization, and the module 'L2N' stands for L2 normalization.*

5.1.2 Architecture

The overall network architecture of the SDL is illustrated in Fig 5.2. We randomly sample $N-$RGB and $N-$IR images, thus resulting in a batch of $2N$ input images. The shaded portion represents the baseline network which outputs the shared representations $\boldsymbol{x}_{\mathcal{RGB}} \in R^d$ and $\boldsymbol{x}_{\mathcal{IR}} \in R^d$. $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ are trained by using the baseline loss $\mathcal{L}_{\mathcal{B}}$. We augment the network with three fully-connected layers represented by $FC3$, $FC4$ and $FC5$ respectively.

The batch normalization outputs of two $FC1$ layers are concatenated together and fed as an input to $FC4$ layer. We call this as spectrum dispelling branch. $FC4$ layer is connected to an identity classifier. The goal of $FC4$ is to learn spectrum-disentangled features and identity discriminative features $\boldsymbol{u}$. $FC3$ and $FC5$ are also trained using disentanglement loss. The learning objective of disentanglement loss is to decompose $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ into $\boldsymbol{v}_{\mathcal{RGB}}$, $\boldsymbol{v}_{\mathcal{IR}}$ and $\boldsymbol{u}$.

We list the outputs and losses for each branch in Table 5.1. We will explain all the terms of Table 5.1 in the subsequent sections.

Table 5.1: *The outputs and losses for each branch of the architecture.*

Branch	Output	Branch Loss	Shared Loss
Baseline	$\boldsymbol{x}_{\mathcal{RGB}}/\boldsymbol{x}_{\mathcal{IR}}$	$\mathcal{L}_{\mathcal{B}}$	
Spectrum dispelling	$\boldsymbol{u}$	$\mathcal{L}_{\mathcal{P}}$	$\mathcal{L}_{\mathcal{D}}$ ($\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}}$ / $\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}}$)
Spectrum distilling	$\boldsymbol{v}_{\mathcal{RGB}}/\boldsymbol{v}_{\mathcal{IR}}$	$\mathcal{L}_{\mathcal{S}}$ ($\mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}}$ / $\mathcal{L}_{\mathcal{S}}^{\mathcal{IR}}$)	

5.1.3 Spectrum Dispelling Branch

The goal of this branch is to learn the identity discriminative and spectrum-disentangled features ($\boldsymbol{u}$) which can efficiently perform RGB-IR Re-ID. To achieve the goal, the spectrum dispelling branch is optimized by the identification loss $\mathcal{L}_{\mathcal{P}}$ and the disentanglement loss $\mathcal{L}_{\mathcal{S}}$ ($\mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}}$ / $\mathcal{L}_{\mathcal{S}}^{\mathcal{IR}}$) together. Since $\boldsymbol{u}$ must capture information relevant to person identity, we use the identification loss and person identity information to supervise the training process. The identification loss $\mathcal{L}_{\mathcal{P}}$ is written as,

$$\mathcal{L}_{\mathcal{P}} = -\frac{1}{N}\sum_{j=1}^{N}\log \boldsymbol{y}_{\mathcal{P}}^{j}, \tag{5.5}$$

where N RGB images and N IR images make up the mini-batch $\mathcal{B}$ and $\boldsymbol{y}_{\mathcal{P}}$ is the predicted probability of the j^{th} input belonging to the ground-truth class. While $\mathcal{L}_{\mathcal{P}}$ is responsible for learning identity discriminative features, the spectrum dispelling effect is achieved by the joint optimization with the disentanglement loss. We explain the disentanglement loss in detail in Section 5.1.5.

5.1.4 Spectrum Distilling Branch

The goal of this branch is to learn spectrum-related information and remove identity related information. For the RGB spectrum feature $\boldsymbol{v}_{\mathcal{RGB}}$ and the IR spectrum feature $\boldsymbol{v}_{\mathcal{IR}}$, we need to train the spectrum distilling branch to fool the identity classifier. Inspired by [105], we first define the so-called "ground truth" identity distribution which is constant over all identities and equal to $1/N_{ID}$, where N_{ID} is the number of identities in the training step. We use the same classifier of $\boldsymbol{u}$ with parameter $\theta_{\mathcal{P}}$ to predict the identity distribution for the RGB spectrum feature $\boldsymbol{v}_{\mathcal{RGB}}$ and the IR spectrum feature $\boldsymbol{v}_{\mathcal{IR}}$. Since the network should embed all identity related information and disentangle it from spectrum component, all the spectrum information needs to be captured in this branch. If the features contain only spectrum related information, the output of the classifier will be confused and the identity-dispeller can easily be fooled by spectrum distilling branch. Note that though the other challenges like pose, illumination and viewpoint variation are present, these challenges are subdued in the baseline features and the major challenge may be attributed to the spectrum.

The identity-dispeller loss combines the identity-dispeller loss for RGB $\mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}}$ and the identity-dispeller loss for IR $\mathcal{L}_{\mathcal{S}}^{\mathcal{IR}}$ together, and it can be written as,

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}} + \mathcal{L}_{\mathcal{S}}^{\mathcal{IR}}, \tag{5.6}$$

where $\mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}}$ and $\mathcal{L}_{\mathcal{S}}^{\mathcal{IR}}$ are respectively

$$\mathcal{L}_{\mathcal{S}}^{\mathcal{RGB}} = \frac{1}{N \times N_{ID}}\sum_{j=1}^{N}\sum_{k=1}^{N_{ID}}(D(\boldsymbol{v}_{\mathcal{RGB}}, \theta_{\mathcal{P}})_k^j - \frac{1}{N_{ID}})^2 \tag{5.7}$$

and

$$\mathcal{L}_{\mathcal{S}}^{\mathcal{IR}} = \frac{1}{N \times N_{ID}}\sum_{j=1}^{N}\sum_{k=1}^{N_{ID}}(D(\boldsymbol{v}_{\mathcal{IR}}, \theta_{\mathcal{P}})_k^j - \frac{1}{N_{ID}})^2. \tag{5.8}$$

For the formulations above, in the mini-batch $\mathcal{B}$, $2N$ images are fed to the network, *i.e.*, N RGB images and N IR images. We set $D(\boldsymbol{v}_{\mathcal{RGB}}, \theta_{\mathcal{P}})_k = \text{softmax}(\mathbf{w}_{\mathcal{P}}\boldsymbol{v}_{\mathcal{RGB}} + \mathbf{b}_{\mathcal{P}})_k$ and $D(\boldsymbol{v}_{\mathcal{IR}}, \theta_{\mathcal{P}})_k = \text{softmax}(\mathbf{w}_{\mathcal{P}}\boldsymbol{v}_{\mathcal{IR}} + \mathbf{b}_{\mathcal{P}})_k$, where $D(\cdot)$ represents the probability distribution over k different identities, $\mathbf{w}_{\mathcal{P}}$ and $\mathbf{b}_{\mathcal{P}}$ refer to the learned weights and biases for $FC4$ layer respectively.

5.1.5 Spectrum Disentanglement

In ideal situation, the spectrum-disentangled feature $\boldsymbol{u}$ contains only person specific cues, while the RGB spectrum feature $\boldsymbol{v}_{\mathcal{RGB}}$ and the IR spectrum feature $\boldsymbol{v}_{\mathcal{IR}}$ contain primarily spectrum information. When we combine the spectrum-disentangled feature and the spectrum feature, it should reconstruct the original feature. The relationship is given by,

$$\boldsymbol{x}_{\mathcal{RGB}} = \boldsymbol{u} + \boldsymbol{v}_{\mathcal{RGB}} \tag{5.9}$$

$$\boldsymbol{x}_{\mathcal{IR}} = \boldsymbol{u} + \boldsymbol{v}_{\mathcal{IR}} \tag{5.10}$$

where $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ are baseline features, $\boldsymbol{u}$ is identity related features and $\boldsymbol{v}_{\mathcal{RGB}}, \boldsymbol{v}_{\mathcal{IR}}$ are spectrum related features. Note that while we use an addition operator, other operators such as Hadamard product can also be used. However, we observe that addition gives better results.

Inspired by this idea, we design the disentanglement loss. If we can learn better spectrum features, the complementry spectrum-disentangled feature will also be learnt better. The disentanglement loss combines the loss $\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}}$ for RGB and the loss $\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}}$ for IR together, and it can be written as,

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}} + \mathcal{L}_{\mathcal{D}}^{\mathcal{IR}}, \tag{5.11}$$

where $\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}}$ and $\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}}$ are respectively

$$\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}} = \frac{1}{N}\sum_{j=1}^{N} ||\boldsymbol{x}_{\mathcal{RGB}}^{j} - (\boldsymbol{u}^{j} + \boldsymbol{v}_{\mathcal{RGB}}^{j})||_2^2 \tag{5.12}$$

and

$$\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}} = \frac{1}{N}\sum_{j=1}^{N} ||\boldsymbol{x}_{\mathcal{IR}}^{j} - (\boldsymbol{u}^{j} + \boldsymbol{v}_{\mathcal{IR}}^{j})||_2^2, \tag{5.13}$$

where $||.||_2^2$ denotes L2 norm. Here, the trivial solution where ($\boldsymbol{v}_{\mathcal{RGB}}$ and/or $\boldsymbol{v}_{\mathcal{IR}}$ can be zero) may happen, and in this sense our method does not guarantee that the spectrum dependent components are always distilled, but our method guarantees that identity relevant information shared between RGB and IR will be kept in $\boldsymbol{u}$.

On the other hand, if we use Hadamard product then,

$$\mathcal{L}_{\mathcal{D}}^{\star} = \mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}\star} + \mathcal{L}_{\mathcal{D}}^{\mathcal{IR}\star} \tag{5.14}$$

where $\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}\star}$ and $\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}\star}$ are respectively

$$\mathcal{L}_{\mathcal{D}}^{\mathcal{RGB}\star} = \frac{1}{N}\sum_{j=1}^{N} ||\boldsymbol{x}_{\mathcal{RGB}}^{j} - (\boldsymbol{u}^{j} \odot \boldsymbol{v}_{\mathcal{RGB}}^{j})||_2^2 \tag{5.15}$$

and

$$\mathcal{L}_{\mathcal{D}}^{\mathcal{IR}\star} = \frac{1}{N}\sum_{j=1}^{N} ||\boldsymbol{x}_{\mathcal{IR}}^{j} - (\boldsymbol{u}^{j} \odot \boldsymbol{v}_{\mathcal{IR}}^{j})||_2^2 \tag{5.16}$$

Note that our spectrum dispelling effect comes from three kinds of losses, *i.e.*, the cross-entropy loss (identification loss), the disentanglement loss, and the identity-dispeller loss. We exploit the disentanglement loss to separate the shared sub-space feature $\boldsymbol{x}$ into spectrum related cues $\boldsymbol{v}$ and identity related cues $\boldsymbol{u}$. Further, we use an identification loss to learn spectrum-invariant or identity related cues $\boldsymbol{u}$, and enhance the identity related information on spectrum-dispelling branch. Additionally, we have an identity-dispeller loss on spectrum distilling branch to extract non-identity and spectrum related features $\boldsymbol{v}$. This has a complementary effect on dispelling branch and promotes identity-specific information and spectrum dispelling effect on $\boldsymbol{u}$. Thus, spectrum dispelling effect is prominent only when all loss functions are applied.

5.1.6 Training Process

The training process has four phases: baseline network training, spectrum dispelling training, spectrum distilling training and joint training. The details are as follows.

Baseline network training. We follow the training procedure of [4] to train the baseline network. The loss function here is $\arg\min \mathcal{L}_{\mathcal{B}}$.

Spectrum dispelling training. We freeze the baseline network and optimize the spectrum dispelling branch, including the fully-connected layer $FC4$ and classifier. The learning rate is set to 0.01 and reduced to 0.1 of its previous value every 60 epochs. The objective function is $\arg\min \mathcal{L}_{\mathcal{P}}$.

Spectrum distilling training. We freeze the baseline network and the spectrum dispelling branch, then we optimize the spectrum distilling branch, including the fully-connected layers $FC3$ and $FC5$. The learning rate is set to 0.01 and reduced to 0.1 of its previous value every 60 epochs. The objective function is $\arg\min \mathcal{L}_{\mathcal{S}}$.

Joint training. We jointly train the entire network in an end-to-end manner and the overall objective function is expressed as

$$\mathcal{L}_{\mathcal{B}} + \lambda_1 \mathcal{L}_{\mathcal{P}} + \lambda_2 \mathcal{L}_{\mathcal{S}} + \lambda_3 \mathcal{L}_{\mathcal{D}}\,. \tag{5.17}$$

The parameters are experimentally tuned to $\lambda_1 = 1$, $\lambda_2 = 10^{-2}$ and $\lambda_3 = 10^{-2}$. Here, we can say that spectrum distilling branch retains the spectrum information, thus $\boldsymbol{u}$ which is obtained from training using identification loss not only preserves identity specific information between the two modalities but is also free of the spectrum component. The overall spectrum disentanglement process gifts the network more ability of disentangling spectrum features. Thus, a joint learning will better balance the network's ability of re-identification and disentanglement.

Table 5.2: *Ablation Study on SYSU-MM01 [5] dataset under indoor-search and all-search mode. CMC-1, CMC-10, CMC-20 (%) and mAP (%) are reported. × refers to the loss function is not used. ✓ represents the applied loss function. '–' means that the parameter is not required. Note that for the variant ① we take a different kind of loss function (hadamard product) of $\mathcal{L}_{\mathcal{D}}$ and discriminate it using mark $\star$.*

	Settings						Indoor Search		All Search	
	$\mathcal{L}_{\mathcal{B}}$	$\mathcal{L}_{\mathcal{P}}$	$\mathcal{L}_{\mathcal{S}}$	$\mathcal{L}_{\mathcal{D}}$	λ_2	λ_3	CMC-1	mAP	CMC-1	mAP
①	×	×	×	✓$\star$	–	–	12.24	22.39	7.44	12.37
②	×	×	×	✓	–	–	20.04	31.12	15.46	19.79
③	✓	✓	×	✓	–	10^{-2}	24.75	36.05	22.53	25.54
④	✓	✓	✓	✓	1	10^{-2}	22.70	30.21	19.22	22.10
⑤	✓	✓	✓	✓	10^{-2}	10^{-2}	28.02	38.06	25.02	26.90

5.2 Experiments

In this section, we describe the datasets and the evaluation protocol. Further, we discuss ablation study to evaluate the contribution of each component of the proposed method. Furthermore, we compare our proposed method with the state-of-the-art methods.

5.2.1 Dataset and Evaluation Protocol

We evaluatae our model on two publicly available datasets SYSU-MM01 [5] and RegDB [6]. The standard Cumulative Matching Characteristics (CMC) value and mean Average Precision (mAP) are adopted [8] to indicate the performance. The testing accuracy is computed using $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ for baseline and $\boldsymbol{u}$ for our network.

SYSU-MM01 [5] is a large-scale dataset and contains 491 persons captured in indoor and outdoor environment. It is collected by six cameras comprising of four RGB and two IR cameras. The training set contains 395 persons, with 22258 visible images and 11909 infrared images. The testing set contains 96 persons. There are two search modes: all-search mode and indoor-search mode. In all-search mode, RGB cameras 1, 2, 4 and 5 are for gallery set and IR cameras 3 and 6 are for probe set. In indoor-search mode, RGB cameras 1 and 2 (excluding outdoor cameras 4 and 5) are for gallery set and IR cameras 3 and 6 are for probe. We follow two evaluation protocols [5] and [4] to evaluate indoor and all-search mode: 1) In the first evaluation protocol, probe set is fixed to 3803 IR images and gallery set has random splits. The overall results are averaged over 10 times. 2) In the second evaluation protocol, probe set is fixed and gallery is formed using 301 randomly selected images. The overall results are reported after averaging 10 times.

RegDB [6] is collected from two cameras and it contains 412 persons. For each person, ten RGB and ten IR images are obtained, respectively. We follow the official protocol [4], where we randomly split the dataset into two halves for training and testing. During testing, images from one modality are used as the gallery set while the ones from the other modality as the probe set.

5.2.2 Implementation details

The model is implemented on PyTorch with NVIDIA-Tesla P100 GPU. Data augmentation includes random cropping and flipping. Input images are resized and random cropped to the size of 227×227 and fed to the network. We use the ResNet-50 [142] architecture as backbone network. The number of epochs is set to 300, and the mini-batch size is 64. Dropout is set to 0.5. We use Adam optimizer for optimizing the overall objective function. The initial learning rate is set to 10^{-2}. It is decreased to 0.1 of its previous value every 60 epochs. All the feature vectors including $\boldsymbol{x}_{\mathcal{RGB}}$, $\boldsymbol{x}_{\mathcal{IR}}$, $\boldsymbol{v}_{\mathcal{RGB}}$, $\boldsymbol{v}_{\mathcal{IR}}$ and $\boldsymbol{u}$ are of dimension 512×1. The source code is available at Github.

Table 5.3: *Ablation Study on parameters (λ_2 and λ_3) for SYSU-MM01 [5] dataset. CMC-1 (%) are reported. The evaluation is for all search mode.*

λ_2	0	1	0	1	1	10^{-3}	2×10^{-4}	5×10^{-4}	10^{-5}
λ_3	0	0	1	1	10^{-3}	1	2×10^{-4}	2×10^{-4}	10^{-5}
CMC-1	16.10	18.36	21.03	18.66	19.16	21.77	23.59	22.16	18.03

Table 5.4: *Comparison with the state-of-the-art methods on SYSU-MM01 [5] datasets under indoor-search and all-search mode. CMC-1, CMC-10, CMC-20 (%) and mAP (%) are reported. $*$ means the evaluation protocol of fixed probe and 301 randomly selected gallery images. # represents the evaluation with fixed probe and random splits of gallery. '–' means that the results are unavailable.*

Approach	**Indoor-search**			**All-search**		
	CMC-1	CMC-20	mAP	CMC-1	CMC-20	mAP
LOMO [138] #	3.89	48.16	8.37	1.75	26.63	3.48
MLBP [152] #	–	–	–	2.12	28.32	3.86
GSM [153]#	–	–	–	5.29	52.95	8.00
Asymmetric FC layer [5]#	14.59	78.68	20.33	9.30	60.38	10.82
One-stream [5]#	16.94	82.10	22.95	12.04	66.74	13.67
Two-stream [5]#	15.60	81.02	21.49	11.65	65.50	12.85
Zero-Padding [5]#	20.58	85.79	26.92	14.80	71.33	15.95
TONE [29]*	–	–	–	12.52	68.60	14.42
HCML [29]*	–	–	–	14.32	69.17	16.16
*cm*GAN [117]#	31.63	89.18	37.00	26.97	80.56	27.80
Kang.*et al.* [154]#	–	–	–	23.18	–	22.49
Baseline [4]*	26.11	85.45	33.21	22.24	79.67	22.10
SDL*	28.02	88.94	38.06	25.02	83.22	26.90
SDL #	**32.56**	**90.67**	**39.56**	**28.12**	**83.67**	**29.01**

5.2.3 Ablation Study

In this subsection, we evaluate the proposed network under different variants. We select the indoor and all-search modes of SYSU-MM01 [5] to do the ablation study. We compare all the different variants in Table 5.2.

Variant ①: We fix the baseline network, and train the whole network with only disentanglement loss $\mathcal{L}_{\mathcal{D}}^{\star}$, which attempts to decompose shared features $\boldsymbol{x}_{\mathcal{RGB}}$ into $\boldsymbol{u}$ and $\boldsymbol{v}_{\mathcal{RGB}}$, and $\boldsymbol{x}_{\mathcal{IR}}$ into

$\boldsymbol{u}$ and $\boldsymbol{v}_{\mathcal{IR}}$. We observe a CMC-1 accuracy of 7.44% for all-search mode.

Variant ②: We keep the baseline network fixed and decompose the baseline features using equations 5 and 6, respectively. This is trained using disentanglement loss given by $\mathcal{L}_{\mathcal{D}}$, described in Section 5.1.5. This loss gives a boost of 8.02% over variant ①. Thus it is evident that the addition operation is more effective than the Hadamard product operation for disentanglement.

Variant ③: We unfreeze the baseline and train the whole system in an end-to-end manner. The loss function for this variant is,

$$\mathcal{L}_{\mathcal{B}} + \lambda_1 \mathcal{L}_{\mathcal{P}} + \lambda_3 \mathcal{L}_{\mathcal{D}} \tag{5.18}$$

where $\lambda_1 = 1$ and $\lambda_3 = 1 \times 10^{-2}$. Here, we observe an accuracy of 22.53% at CMC-1. Thus the joint loss of $\mathcal{L}_{\mathcal{P}}$, $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{B}}$ helps to learn the identity cues at the spectrum dispelling branch, while filtering out the spectrum related information at the spectrum distilling branch and increases the accuracy significantly.

Variant ④:. We augment the network with an identity-dispeller to fool the identity classifier and remove the spectrum related information at spectrum dispelling branch. The overall system is optimized with the objective function in Eq. 6.14 by setting $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 10^{-2}$. However, a large weight given to the identity-dispeller loss adversely influences the learned weights of the classifier which makes the network less effective for Re-ID task. Thus we observe that the accuracy reduces to 19.22%.

Variant ⑤:. To alleviate the problem of variant ④ , we give less weightage to the identity-dispeller loss. Here, we use a weighted identity-dispeller with unfreezed baseline and observe the best performance of 25.02% at CMC-1. The overall objective function is Eq. 6.14, and we set $\lambda_1 = 1$, $\lambda_2 = 10^{-2}$ and $\lambda_3 = 10^{-2}$. It demonstrates that the SDL network efficiently disentangles the spectrum component from the features. Thus there is a trade-off between how much spectrum information is distilled at spectrum distlling branch while affecting the identity discriminative ability at spectrum dispelling branch.

In case of the Indoor-search mode, when we take $\mathcal{L}_{\mathcal{B}}$, $\mathcal{L}_{\mathcal{P}}$, $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{D}}$, the accuracy is 22.70% with λ_2 =1 and λ_3 = 0.01 for Variant ④. Here, the accuracy is lower than the baseline accuracy of 26.11%. The reason is as follows. $\mathcal{L}_{\mathcal{S}}$ loss is applied to achieve the goal of spectrum distillation. The spectrum distillation branch is trained so as to fool the identity classifier thereby distilling only the spectrum related features. With a significantly higher weight λ_2 associated with $\mathcal{L}_{\mathcal{S}}$, this loss may dominate other losses. Since the primary aim of $\mathcal{L}_{\mathcal{S}}$ is to fool the identity classfier, this results in poor Re-ID performance. However, with a balanced weight λ_2, the loss $\mathcal{L}_{\mathcal{S}}$ promotes learning spectrum features without dominating other losses and we observe a better accuracy of 28.02% for for Variant ⑤ which is an improvement by 1.91% over baseline network features for indoor-search.

We report additional ablation study for all-search mode of SYSU-MM01 [5] to see the behaviour of different combination of loss functions and the effect of loss parameters in Table 5.3. Here, we observe that disentanglement loss $\mathcal{L}_{\mathcal{D}}$ gives better accuracy than identity-dispeller loss $\mathcal{L}_{\mathcal{S}}$. We also fixed $\lambda_1 = 1$ and changed the values of λ_2 and λ_3. We observe that if we give less weight to λ_2 and λ_3, they nullify the effect of disentanglement ($\mathcal{L}_{\mathcal{D}}$) and identity-

dispeller loss ($\mathcal{L}_{\mathcal{S}}$). Further, if both are weighted between 10^{-2} to 10^{-4}, then it gives a high accuracy. It may be noted that when we use $\lambda_2 = \lambda_3 = 0$, *i.e.*, in the absence of $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{D}}$, the network is not well trained and $\boldsymbol{u}$ is not learnt completely. In Table 5.3, the accuracy of 16.10% is reported using this $\boldsymbol{u}$. Whereas, in Table 5.4, baseline network features $\boldsymbol{x}_{\mathcal{RGB}}$ and $\boldsymbol{x}_{\mathcal{IR}}$ are used. Since the baseline is already well trained, the CMC-1 accuracy is higher.

5.2.4 Comparison with State-of-the-art

In this section, we compare our proposed approach against hand-crafted feature learning methods including LOMO [138], MLBP [152] and GSM [153] and various deep learning based methods like Asymmetric FC layer [5], One-Stream [5], Two-stream [5], Zero padding [5], TONE [29], *cm*GAN [117] and baseline [4].

Results on SYSU-MM01 [5]

We report the CMC rank accuracies for SYSU-MM01 [5] dataset in Table 5.4 under indoor-search and all-search mode. Among them, all-search mode is more challenging. Two types of evaluation settings are available in the literature. To make a fair comparison, we report the results for both settings.

Evaluation-I: It is represented as '$*$' in Table 5.4. [4, 29] follow the setting of fixed probe and randomly select 301 gallery images. Based upon this evaluation, we observe an increase of 10.22% over the best network of [29] and 2.78% over baseline [4] [1] under all-search mode for CMC-1. The superiority of the proposed method is attributed to the design of spectrum-disentangled representations.

Evaluation-II: It is represented as '#' in Table 5.4. [5, 117, 154] follow the setting of fixed probe with random splits of gallery. Here, we observe an accuracy of 28.12% and mAP of 29.01% under all-search mode for CMC-1. In addition, our model also converges quickly within 300 epochs. On the other hand, state-of-the-art methods like *cm*GAN [117] apply adversarial training which attains convergence after 2800 epoches.

Results on RegDB [6]

We report the CMC rank accuracies for RegDB [6] dataset in Table 5.5. We evaluate the performance on different query settings: RGB to IR and IR to RGB. RGB to IR means that probe set relates to RGB images and gallery set to IR images during testing. Similarly, IR to RGB corresponds to probe to gallery matching. IR to RGB setting is difficult and challenging for RegDB dataset as compared to SYSU-MM01 due to difference in wavelength.

RGB to IR: We achieve an accuracy of 26.47% for RGB to IR setting.

IR to RGB: We observe an accuracy of 25.74% at CMC-1.

The difference in these results is marginal which bolsters our understanding of the network

[1] Please note that the baseline results is reproduced using PyTorch version [155].

that the representations do not carry any significant spectrum information. We can also observe that the proposed method outperforms the competing methods [5, 29, 138] and baseline [4] consistently by a huge margin on both settings.

Table 5.5: *Comparison with the state-of-the-art methods on RegDB [6] for different query settings. CMC-1, CMC-10, CMC-20 (%) and mAP (%) are reported.*

Approach	RGB to IR			
	CMC-1	CMC-10	CMC-20	mAP
LOMO [138]	0.85	2.47	4.10	2.28
MLBP [152]	2.02	7.33	10.90	6.77
GSM [153]	17.28	34.47	45.26	15.06
One-stream [5]	13.11	32.98	42.51	14.02
Two-stream [5]	12.43	30.36	40.96	13.42
Zero-Padding [5]	17.75	34.21	44.35	18.90
TONE [29]	16.87	34.03	44.10	14.92
TONE + HCML [29]	24.44	47.53	56.78	20.08
Baseline [4]	22.02	48.78	57.44	21.45
SDL	**26.47**	**51.34**	**61.22**	**23.58**
Approach	**IR to RGB**			
	CMC-1	CMC-10	CMC-20	mAP
Zero-Padding [5]	16.63	34.68	44.25	17.82
TONE [29]	13.86	30.08	40.05	16.98
TONE + HCML [29]	21.70	45.02	55.58	22.24
Baseline [4]	20.34	47.22	55.09	21.11
SDL	**25.74**	**50.23**	**59.66**	**22.89**

5.2.5 Visualization of Person Features through t-SNE

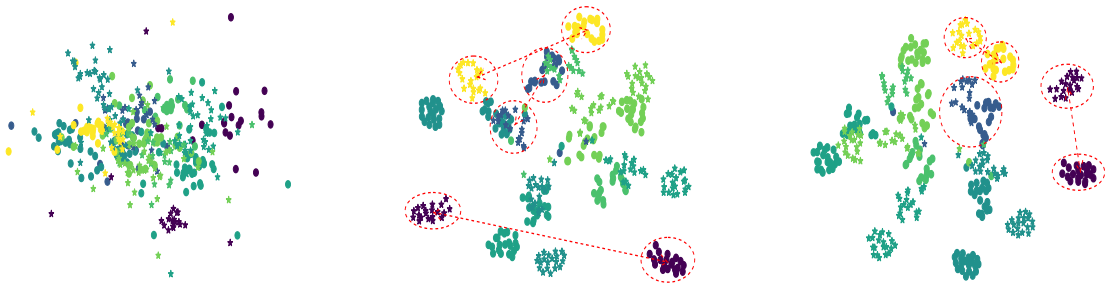
We randomly select 300 IR and RGB images of 8 identities from the testing set respectively and visualize the features using t-SNE [140] in 6.4. Here, the shape ★ represents the person features generated by IR images and the shape ● represents the person features generated by RGB images. Features corresponding to the same identity are shown by the same color. We show features for three phases: initial phase (without training), baseline features and spectrum-disentangled features from the proposed approach in 6.4 (from left to right).

Initial phase : As expected, we can observe that all the data points are dispersed and some are totally merged into samples of another identity.

Baseline phase: In 6.4(b), we observe that the features obtained from the trained baseline cluster very well. Thus, we note that the baseline feature $\boldsymbol{x}$ is robust to pose, viewpoint variations *etc*. However, the cross-modality features are far apart due to large spectrum difference. Therefore, once $\boldsymbol{x}$ is learnt, the major challenge is to disregard the spectrum information. For example, samples in purple and yellow color are still farther apart.

SDL phase: Since the baseline phase does not disregard the spectrum information, the SDL phase addresses this challenge by removing the spectrum information. We can observe that

features of same identity but from different modality points appear closer. In particular, the distance between purple, yellow and blue samples has reduced drastically. This shows the superiority of spectrum-disentangled representation over the baseline. Thus, we conclude that the network removes spectrum related information and maintains useful cues for a person' images belonging to different modality.



(a) Initial Features (b) Baseline Features (c) Disentangled features

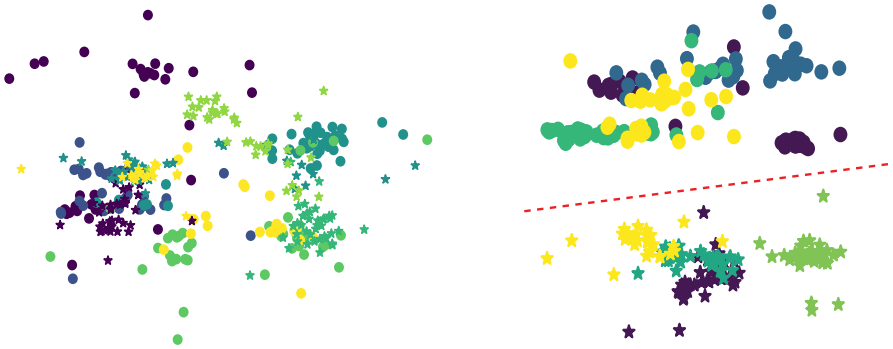
Figure 5.3: 2D visualization of the features. *Different color stands for different persons. The shape ★ denotes the feature extracted from the IR image, while the shape ● denotes the feature extracted from the RGB image. For better clarity, please zoom-in at 400%.*

5.2.6 Visualization of Spectrum Features through t-SNE

In order to understand the disentaglement effect of the network, we conduct experiment with RGB-RGB and RGB-IR images where we give RGB-RGB and RGB-IR images as an input to the SDL network. Further, we extract features from spectrum distilling branches $FC3$ and $FC5$ and visualize t-SNE plots.

In the first case, we input only RGB images to both the branches and finetune the network with RGB images only. Further, we extract features from spectrum-distilling branch. We visualize the features in 5.4(a). Here, different colors refer to different identities, and different shapes refer to features obtained from $FC3$ and $FC5$. We can see that the points do not form any visible clusters on the basis of spectrum. This is expected as we have only RGB spectrum input.

In the second case, we use RGB-IR images as input to the two branches. We visualize the spectrum features extracted at spectrum distilling branch in 5.4(b). Here, we can observe that spectrum features generated from RGB images are quite distinct from features of IR images. Further, the features of different identities (different colors) do not show any discriminative nature. Thus, it is clearly evident that the network has strong ability to distill spectrum and dispel identity at spectrum distilling branch.



(a) RGB-RGB (b) RGB-IR

Figure 5.4: 2D visualization of the spectrum features. *Different color stands for different persons. The shape ★ denotes the feature extracted from the upper spectrum distilling branch, FC3, while the shape ● denotes the feature extracted from the lower spectrum distilling branch, FC5.*

Table 5.6: *Results with Data Augmentation (Random-Erasing at three levels) on SYSU-MM01 datasets under all-search mode. CMC-1, CMC-10, CMC-20 (%) and mAP (%) are reported.*

Random-Erasing	All-search			
	CMC-1	CMC-10	CMC-20	mAP
0%	**28.12**	**70.23**	**83.67**	**29.01**
5%	24.75	69.52	83.66	28.81
10%	23.86	68.13	83.32	27.41
20%	23.42	67.90	82.90	27.30

5.2.7 Effect of Data Augmentation

We apply random erasing as a data augmentation method on the SYSU-MM01 dataset. We randomly erase a patch with 5%, 10%, and 20% area of the training image. Table 5.6 reports the results. It shows that the performance decreases as the area of random erasing increases. We also observe that the training time increases in proportion to that of the area of random erasing, since the erased patch will make training hard. Hence, we can conclude that random erasing is not a good technique for the visible-infrared person re-identification task.

5.2.8 Retrieval Results

We visualize retrieval results on SYSU-MM01 [5] in 5.5 using the baseline [4] and the proposed SDL network on same query images. We randomly pick IR images from the probe set as query and retrieve the corresponding RGB images from the gallery set. The images in the first column are the query images. The retrieved images are sorted according to the similarity scores from left to right (from second column till last). We can see that the proposed SDL network gets good shots under most of the situations. Similarly, we randomly select some RGB images from probe to match against IR images on RegDB [6] dataset. These results are shown in 5.6.

(a) Top-5 results from Baseline [4] (b) Top-5 results from SDL Network

Figure 5.5: *Top-5 retrieval results on SYSU-MM01 [5] dataset under all-search mode. Red boundary indicates a negative match and green shows a positive match.*

(a) Top-5 results from basline [4] (b) Top-5 results from SDL Network

Figure 5.6: *Top-5 retrieval results on RegDB [6] dataset. Red boundary indicates a negative match and green shows a positive match.*

This chapter covers the challenging scenerio of RGB-IR problem. Towards this, we discuss a disentanglement network to address the spectrum challenge. In the next chapter, we discuss text-image matching problem.

Chapter 6

Text-Image based Re-ID

In this chapter, we describe our proposed framework towards the emerging area of description based Re-ID task for heterogeneous Re-ID. We discuss the major challenges in text-image Re-ID in section 6.1. Section 6.2 covers the proposed hierarchical attention image-text alignment network (HAITA-Net). We evaluate our algorithm with different variants in section 6.3. We also report results of comparison with other popular Re-ID algorithms.

6.1 Text Analysis

In particular, we address two major challenges of the description based Re-ID task which is ignored in the past. Our first observation is that there are certain words which do not match with any image region. Such text-image non-correspondence causes the challenge of alignment uncertainty and deteriorates the matching capacity of the network. Thus, a correct alignment mechanism of text description and image regions is necessary for efficient Re-ID. Second, we observe a significant text complexity due to the fact that the query descriptions can widely differ for the same image while having the same semantics because of different annotators. For example, annotators can describe a dress in either *"a light dress" or "a white dress"*. This statement states two critical observations: i) words can vary a lot across the annotations for the same image; ii) some words can be more important than other words having same meaning like *white* is more informative than *light*. This problem is also known as the word-variance problem. Thus in this case, the network has to try hard to map all these different words to the same semantics.

6.2 Proposed HA2-Net

Our model is motivated by the fact that effective matching can be obtained by building a strong similarity between salient words and image patches so that the network can efficiently navigate the visual content via textual information. The architecture of the proposed model HAITA-Net is shown in Figure 6.1. The architecture consists of a baseline network for associating the text-image features (the dashed boxes of Figure 6.1), an efficient technique of Threshold Term Frequency Inverse Document Frequency (T2FIDF) to find out the most salient tokens, and an attention mechanism to build a hierarchical relationship between text and image (the green shaded portion in Figure 6.1).

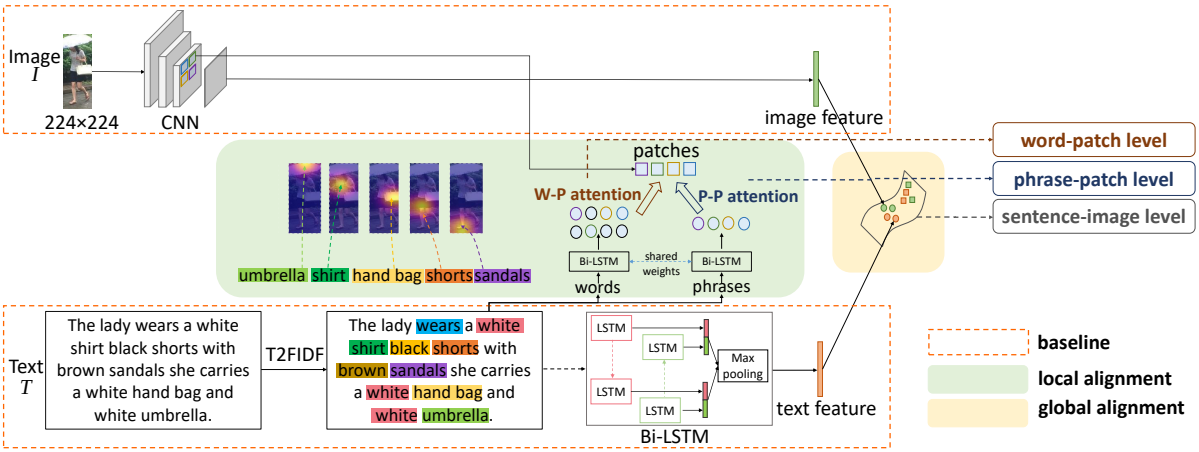


Figure 6.1: *The architecture of HAITA-Net. HAITA-Net consists of a global alignment part and a local alignment part. For the global alignment part, CNN and Bi-LSTM models encode the image and text features respectively, and they are projected to the global alignment subspace to build the relationship at a sentence-image level. For the local alignment part, a hierarchical attention alignment component is proposed, including word-patch level and phrase-patch level. In particular, T2FIDF is applied to descriptions to find out the most salient tokens. The highlighted text represents the most salient tokens and some of the corresponding attentions are shown in the shaded portion of the local alignment part. Best viewed in color*

6.2.1 Baseline

The baseline network consists of a deep convolutional neural network and a bi-directional LSTM (Bi-LSTM) to encode the image and text features. We input an image I of size 224×224 to obtain the image features. On the other hand, we pre-process the textual description T and split it into words, and then sequentially input them to Bi-LSTM. Here, we concatenate the hidden states of forward and backward directions, and text features are obtained with a max-pooling strategy.

The extracted image and text features are projected into a sub-space, where the features are associated and the similarity between the features is maximized. This is depicted in global alignment in Figure 6.1. The baseline network is optimized through cross-modal losses used in cross-modal applications [118, 156, 157] and is represented as $\mathcal{L}_B$ given in Eq.6.1. The baseline loss comprises of the cross-modal projection classification (CMPC) and cross-modal projection matching (CMPM) loss. The CMPC loss attempts to categorize the vector projection of representations from one modality onto another with the improved norm-softmax loss. The CMPM loss minimizes the KL divergence between the normalized matching distributions to associate the representations across different modalities.

$$\mathcal{L}_B = \mathcal{L}_C^I + \mathcal{L}_C^T + \mathcal{L}_M^{I2T} + \mathcal{L}_M^{T2I}, \tag{6.1}$$

where $\mathcal{L}_C^I$, $\mathcal{L}_C^T$, $\mathcal{L}_M^{I2T}$ and $\mathcal{L}_M^{T2I}$ are defined as follows.

$$L_C^I = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(W_i^T\phi(I_i)+b_i\right)}{\sum_j \exp\left(W_j^T\phi(I_i)+b_j\right)}\right) \quad s.t.\ ||W_j|| = 1, \forall j \in \{1,2,...,N\} \tag{6.2}$$

where $||.||$ represents normalization, C denotes classification, I corresponds to the image modality, N is the number of batch size, and W_i, b_i are the weights and the bias of the classification layer for the visual feature representation $\phi(I)$. The same procedure is followed to compute the text classification loss, $\mathcal{L}_C^T$.

$$L_M^{T2I} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right) \tag{6.3}$$

where M stands for matching, $T2I$ refers to the text-to-image matching, ϵ is a very small number for preventing division by zero, N represents batch size, $p_{i,j}$ represents the proportion of the scalar projection of the image embedding and the normalized textual embedding among all scalar projections between pairs in a batch and $q_{i,j}$ represents the normalized true matching probability in each batch. We can also calculate the matching in opposite direction (i.e., from image to text) and compute loss $\mathcal{L}_M^{I2T}$.

6.2.2 Threshold TF-IDF (T2FIDF)

We propose the Threshold Term Frequency Inverse Document Frequency (T2FIDF) technique to find out the most salient tokens from the textual description. The output of T2FIDF is the highlighted words in the box. We observe that less frequent words carry most information which results in more discriminative features. Further, frequent words can be redundant and irrelevant for matching. The main goal of TF-IDF [158] is to give more weight to the less frequent words. To calculate TF-IDF, we consider every caption c in the text as a separate document to build a corpus C. In TF-IDF, the "term-frequency" is simply a count of the number of times a word w appears in a given context, while the "inverse document frequency" term puts a penalty on how often the word appears elsewhere in the corpus. The final score is calculated as

$$tfidf(w,c,C) = f(w,c) * log\frac{N}{|c \in C; w \in c|}, \tag{6.4}$$

where $f(w,c)$ indicates the term frequency, N is the total number of captions, and the denominator represents the number of captions in which the word w appears.

We apply a threshold value to find out the most important tokens. The threshold value is experimentally set to 0.01. All the words having $tfidf$ scores greater than 0.01 are considered as salient words. For example, words like "shirt", "pants" and "umbrella" are salient and most frequent words like "light", "man" and "woman" are discarded.

6.2.3 Hierarchical Attention Alignment

The hierarchical relationship builds upon three levels: word-patch (W-P) level attention, phrase-patch (P-P) level attention, and sentence-image level. Our motivation for building such relationship is to leverage the local salient regions which can produce discriminative features in conjunction with global features. We describe these levels in the following.

Word-Patch (W-P) Level Attention

We process the salient tokens extracted from T2FIDF to get feature vectors. The extracted salient tokens are fed to the Bi-LSTM to encode the word-representations. Then, we use the output of each Bi-LSTM unit and build a relationship between all the salient word features and patches via attention mechanism. The feature matrix of all salient words is indicated by $s_w \in R^{D \times W}$ where s_{wi} is the feature vector for the i^{th} word. D is the dimension of the word vector and W is the number of words. Similarly, we extract the patch feature matrix of size $512 \times 14 \times 14$ from the "$Conv\ dw$" layer of Mobile-Net-v1. Then, we reshape it into $f \in R^{512 \times 196}$. Here, each column of f is the feature vector of a sub-region of the image. 512 is the dimension of the local feature vector, and 196 is the number of patches.

Inspired by [159], we design an attention module to measure the fine-grained matching through the potential relationship between word and patches. To accomplish the task, we first calculate the similarity matrix for all possible salient word features s_w and patch features f given by

$$S = {s_w}^T f, \tag{6.5}$$

where $S \in R^{W \times 196}$ and $S_{i,j}$ is the similarity between the i^{th} word of the sentence and the j^{th} sub-region of the image. Then, we normalize the similarity matrix $S_{i,j}$ by

$$S_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=0}^{W-1} \exp(S_{k,j})}. \tag{6.6}$$

The attention vector is calculated *w.r.t* all the patches to find out the most attentive image region and is given by

$$\alpha_j = \frac{\exp(S_{i,j})}{\sum_{k=0}^{196} \exp(S_{i,k})}. \tag{6.7}$$

We use the above attention vectors α_j to compute the most attentive region vector (β_i) for each word. It is computed as the weighted sum over all regional visual vectors, *i.e.*

$$\beta_i = \sum_{j=0}^{196} \alpha_j f_j. \tag{6.8}$$

Further, we calculate the relevance between each salient word and the attentive image region using the cosine similarity,

$$R(s_{wi}, \beta_i) = \frac{{s_{wi}}^T \beta_i}{||s_{wi}||||\beta_i||}. \tag{6.9}$$

Then we compute the posterior probability of word matching with image region as follows

$$P(s_{wi}|\beta_i) = \frac{\exp(R(s_{wi}, \beta_i))}{\sum_{k=0}^{N} \exp(R(s_{wk}, \beta_i))}, \tag{6.10}$$

where N is the number of batch size.

Here, we minimize the word-patch loss function which is a negative log posterior matching probability given by

$$\mathcal{L}_{\mathcal{W}} = -\sum_{i=1}^{N} \log P(s_{wi}|\beta_i). \tag{6.11}$$

Phrase-Patch (P-P) Level Attention

The phrase-patch level relationship has more precise correspondences with image parts than isolated words. A noun phrase describes a specific region in the image, thus the phrase feature is more related to local visual features. Therefore, we focus on noun phrases and encode them separately[1]. We use the Bi-LSTM to encode the noun phrase and then build the relationship between nouns and patch in a similar manner as we discussed above for word-patch. As described in Eq. 6.5, we first calculate the similarity matrix for all possible phrase features s_p and patch features f given by

$$S^{phrase} = {s_p}^T f, \tag{6.12}$$

where $S^{phrase} \in R^{P \times 196}$ and P is the number of noun-phrases $S^{phrase}_{i,j}$ is the similarity between the i^{th} phrase of the sentence and the j^{th} sub-region of the image. By replacing S with S^{phrase} in Eq. 6.6 and Eq. 6.7, we obtain β using Eq. 6.8. To compute relevance R, we replace s_w by s_p and obtain the loss function $\mathcal{L}_{\mathcal{P}}$ using Eq. 6.10 and Eq. 6.11.

Here, we minimize the phrase-patch loss function which is a negative log posterior matching probability given by

$$\mathcal{L}_{\mathcal{P}} = -\sum_{i=1}^{N} \log P(s_{p_i}|\beta_i). \tag{6.13}$$

Sentence-Image Level

The image and sentence feature vectors are projected on the global sub-space to build the relation between sentence and image. The loss function $\mathcal{L}_B$ is used which is described in Section 3.1.

In our experiments, we demonstrate that the hierarchical attention mechanism gives a significant boost over the baseline.

[1]To extract nouns, we use the Natural Language ToolKit (NLTK) [160].

6.2.4 Training Process

The training process has four phases: baseline network training, W-P attention training, P-P attention training, and joint training. The details are as follows.

Baseline Network Training. The baseline network is trained using baseline loss functions $\mathcal{L}_B$ described in Section 3.1.

W-P Attention Training. We freeze the baseline network and optimize the W-P attention module. The objective function is $\mathcal{L}_W$ given in Eq. 6.11.

P-P Attention Training. We freeze the baseline network and the W-P attention, then we optimize the P-P attention. The objective function is $\mathcal{L}_\mathcal{P}$ given in Eq. 6.13.

Joint Training. We jointly train the entire network in an end-to-end manner and the overall objective function of the hierarchical weighted loss function can be expressed as follows

$$\mathcal{L}_T = \mathcal{L}_B + \lambda_1 \mathcal{L}_W + \lambda_2 \mathcal{L}_P, \tag{6.14}$$

where the parameters of λ_1 and λ_2 are the weights provided to the loss functions and are experimentally tuned. The values of λ_1 and λ_2 are provided in the ablation study part.

6.3 Experiments

In this section, we describe the datasets and the evaluation protocol. Further, we discuss the ablation study to evaluate the contribution of each component of the proposed method, provide a comparison with the state-of-the-art methods, and then make elaborate visualization analysis.

6.3.1 Datasets

We evaluate HAITA-Net on two standard publicly available dataset: CUHK-PEDES [10] and Flickr30K [11].

CUHK-PEDES: The dataset used is CUHK-PEDES, which is currently the only benchmark for description based person Re-ID. It is a collection of five existing person Re-ID datasets, CUHK03 [7], Market-1501 [8], SSM [161], VIPER [162], and CUHK01 [163]. It contains 40,206 pedestrian images of 13,003 identities, with each image described by two textual descriptions. The dataset is split into 11,003 training identities with 34,054 images, 1000 validation persons with 3,078 images and 1000 test individuals with 3,074 images, without having overlaps with same person IDs. On average, each textual description contains more than 23 words. The dataset contains 9408 different words.

Flickr30K: The Flickr30K dataset contains 31,783 images with each image annotated by five text descriptions which contain a wide variety of images (humans, animals, objects, scenes). The data is split into 29,783 images for training, 1,000 images for validation, and 1,000 images for testing.

6.3.2 Evaluation Metrics

The standard Cumulative Matching Characteristics (CMC) value and mean Average Precision (mAP) are adopted [8] to indicate the performance.

6.3.3 Implementation details

The model is implemented with TensorFlow and runs on NVIDIA-Tesla P100 GPU. The training set is enlarged by data augmentation which includes random cropping and flipping. Input images are resized and random cropped to the size of 224×224 and fed to the network. We use the MobileNet architecture as the backbone network and Bi-LSTMs, respectively. The number of training epochs is 50 and the mini-batch size is 16. We use Adam optimizer for optimizing the overall objective function. The initial learning rate is set to 2×10^{-4}. The rate is decreased to 0.1 of its previous value for every 15 epochs. For Flickr30K [11], we report the results with ResNet-152 as the image feature extractor.

6.3.4 Ablation Study

We evaluate the proposed network under different variants. We select the CUHK-PEDES [10] dataset to do the ablation study and compare all the variants in Table 6.1.

Variant ①: We compare different combinations of the sub-networks in Table 6.1. We first experiment with the pre-trained visual networks like VGG-16 [164], ResNet-152 [142], and Mobile-Net [165] as the visual backbone and train them on text-image Re-ID datasets. For the textual domain, we experimented with one hot encoded vector, LSTMs and Bi-LSTMs. We observe that Mobile-Net and Bi-LSTMs give the best performance for CUHK-PEDES and ResNet-152 and Bi-LSTM work well for Flickr-30K. Hence, we keep this as the baseline network to encode the image and text features. The network is trained via loss $\mathcal{L}_{\mathcal{B}}$ [118]. We observe a CMC-1 accuracy of 50.02% for the CUHK-PEDES.

Variant ②: Here we use weighted TFIDF and thresholded TFIDF (T2FIDF) to compute most salient word tokens. In case of weighted TFIDF, we weigh all word vectors by multiplying them with the TFIDF score computed using Eq. 6.4. In case of T2FIDF, a threshold value is decided experimentally to discard the word vectors having less score than threshold. We see a decrease in baseline accuracy when the weighted TFIDF is applied. On the other hand, thresholding discards the uninformative word vectors based on the threshold value and gives a boost of 3.1% over the weighted TFIDF.

Variant ③: In this variant we introduce word-patch attention and the network is trained with $\mathcal{L}_{\mathcal{T}}$ given in Eq. 6.14. In the first case, we set $\lambda_1 = 1$ and $\lambda_2 = 0$. This setting tries to capture attention between all possible word and patch (W-P) pairs. The version Unit in Table 6.1 gives the accuracy for this setting. However, we see a decrease in the Re-ID accuracy compared to the best model of Variant ②. This is due to the fact that label information for every corresponding word and patch is not available and there exist certain words that do not have corresponding regions. To address this issue, we use weighted word-patch attention. We show how accuracy is getting affected with different weight values of λ_1 in Figure 6.2. Among all the

values, we observe that $\lambda_1 = 10^{-6}$ results in an increase of 4.66% in the CMC-1 accuracy over Unit word-patch attention. We further observe that higher value of λ_1 lead to poor accuracy. This is expected as the network tries to focus too much on local attentive regions and will learn to efficiently match local regions as compared to the final goal of global matching. We also incorporated patch to word attention, however, we did not find any improvement.

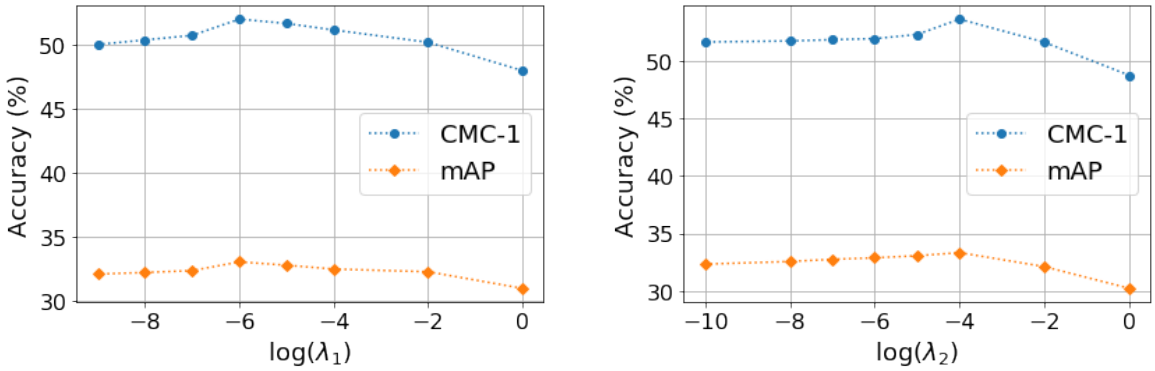


Figure 6.2: *Effects of different values of λ_1 and λ_2.*

Table 6.1: *Ablation Study on the CUHK-PEDES dataset. mAP, CMC-1, and CMC-10 (%) are reported.*

	Variants	**CUHK-PEDES**		
		mAP	CMC-1	CMC-10
① Baseline	VGG-16 +One-hot-encoding	12.02	13.11	42.66
	VGG-16 + LSTM	19.66	25.94	62.44
	VGG-16 + Bi-LSTM	26.66	45.20	76.55
	ResNet-50 + Bi-LSTM	31.39	49.03	80.60
	ResNet-152 + Bi-LSTM	31.64	49.16	80.07
	ResNet-101 + Bi-LSTM	31.85	49.84	80.72
	MobileNet + Bi-LSTM	32.06	50.02	81.34
② TF-IDF	Weighted	31.67	47.34	81.45
	Threshold (T2FIDF)	32.43	50.44	81.37
③ W-P Attention	Unit	30.63	47.76	79.84
	Weighted	33.01	52	81.87
④ P-P Attention	HAITA	33.33	53.67	83.79
	HAITA-Net	**37.08**	**56.57**	**86.30**

Variant ④:. In this variant, we introduce phrase-patch (P-P) based attention and the network is trained with $\mathcal{L}_{\mathcal{T}}$. We set $\lambda_1 = 10^{-6}$ and vary the values of λ_2. We observe that high weightage to λ_2 leads to decrease in the accuracy and the value of $\lambda_2 = 10^{-4}$ works well. We first extract nouns and feed it into Bi-LSTM to get phrase features. Here, we observe a boost of 4.07% in CMC-1 accuracy with re-ranking (RK) [66] over the Variant ③ in case of HAITA-Net.

6.3.5 Comparison

We show the comparison results with the state-of-the-art methods on the Re-ID dataset CUHK-PEDES. We conducted bi-directional matching results: text-to-image and image-to-text Re-ID.

Text-to-Image Re-ID

For CUHK-PEDES, we compare with feature embedding based methods, such as MCCL [119] and CMPM+CMPC [118]. The results are shown in Table 6.2. We also compare with attention based methods, including GNA-RNN [10], CNN-LSTM [127], DSFA-Net [166], PWM + ATH [167], GDA+LRA [129], MIA [168], and A-GANet [121]. The results show that our HAITA-Net achieves better performance. Although MIA [168] and A-GANet [121] also learn the cross-modal representations by global or local associations, they weakly incorporate the importance of the informative word. Further, compared to A-GANet, we observe that for higher CMC ranks our model has significantly better performance. Hence, an improved performance over these methods suggests that our hierarchical attention network with salient tokens can learn more discriminative and robust representations.

Table 6.2: *Comparisons on the CUHK-PEDES dataset (Text-to-image Re-ID). CMC-1 and CMC-10 (%) are reported. A:Embedding and B:Attention*

	Method	Baseline	Param	Ref	CUHK-PEDES	
					CMC-1	CMC-10
A	CMPM+CMPC [118]	MobileNet+Bi-LSTM	6M	ECCV'18	49.37	79.27
	MCCL [119]	MobileNet+Bi-LSTM	6M	ICASSP'19	50.58	79.06
	TIMAM [169]	ResNet-101+BERT	152M	ICCV'19	54.51	84.78
	CCA [170]	ResNet-50+Text-CNN	25M	Arxiv'20	46.44	76.30
	Dual-Path [171]	ResNet-50+Text-CNN	25M	TOMM'20	44.40	75.07
B	GNA-RNN [10]	VGG-16 + One-hot	138M	CVPR'17	19.05	53.64
	CNN-LSTM [127]	VGG-16 + LSTM	138M	ICCV'17	25.94	60.48
	DSFA-Net [166]	VGG-16 + LSTM	138M	PRL'18	20.33	54.90
	PWM+ATH [167]	VGG-16+LSTM	138M	WACV'18	27.14	61.02
	GDA+LRA [129]	ResNet-50 + LSTM	25M	ECCV'18	43.58	76.26
	A-GANet [121]	ResNet-50 + Bi-LSTM	25M	ACM MM'19	53.14	81.95
	Pythia-reID [172]	ResNet-152 + GRU	60M	Robotics'20	26.79	63.78
	MIA [168]	VGG-16 + GRU	138M	TIP'20	48.00	79.30
	MIA [168]	ResNet-50 + GRU	25M	TIP'20	53.10	82.90
	PMA [173]	VGG + Bi-LSTM	138M	AAAI'20	47.02	78.06
	PMA [173]	ResNet-50 + Bi-LSTM	25M	AAAI'20	53.81	81.23
	HAITA-Net	**MobileNet + Bi-LSTM**	**6M**	**Ours**	**56.57**	**86.30**

Image-to-text Re-ID

We also conducted experiments in the image-to-text Re-ID setting. The results for the CUHK-PEDES dataset are shown in Table 6.3. The results show that the HAITA-Net is able to perform good bidirectional matching task as compared to the existing algorithms Bi-rank [124], Histogram [125], N-pair [126], CMPM+CMPC [118], and baseline (see Section 3.1). The results also show that our HAITA-Net outperforms the existing methods.

Table 6.3: *Comparisons on the CUHK-PEDES dataset (Image-to-Text Re-ID). CMC-1, CMC-5, and CMC-10 (%) are reported.*

Method	CUHK-PEDES		
	CMC-1	CMC-5	CMC-10
Bi-rank [124]	32.56	–	–
Histogram [125]	4.78	–	–
N-pair [126]	17.66	–	
CMPM+CMPC [118]	57.71	–	91.28
Baseline	61.61	86.14	92.0
HAITA-Net	**65.23**	**88.11**	**94.13**

6.3.6 Visualization

Attention Visualization through Heatmaps

We select three examples to show the effectiveness of our attention alignment modules in Figure 6.3. For each row, from left to right are the text with a highlighted word, the image, the heatmap for the highlighted word by our W-P attention module, and the heatmap for the highlighted word by both W-P and P-P attention modules, respectively. The lighter regions indicate the highly attentive areas for the corresponding highlighted words.

The figure shows that the Word-Patch (W-P) attention module emphasizes patches corresponding to isolated words. For example, if we search for the "handbag" keyword then W-P attention focuses on the hand. Similarly, in the second case of the keyword "backpack", the attended region is back. In the third case of the keyword "umbrella", the W-P attention focuses on the umbrella of the other person than the target one. On the other hand, HAITA-Net comprises of both Word-Patch and Phrase-Patch (W-P+P-P) and selects the corresponding regions efficiently related to the highlighted word and filters out the unrelated regions. For example, in the case of "handbag", the focus is on handbag location. The similar behavior in other examples illustrates that HAITA-Net learns accurate aligned fine-grained matching.

T-SNE Visualization of Text-Image Embeddings

We randomly select five identities from the testing set of CUHK-PEDES dataset and visualize the features using t-SNE in Figure 6.4. Here, different colors refer to different identities, and

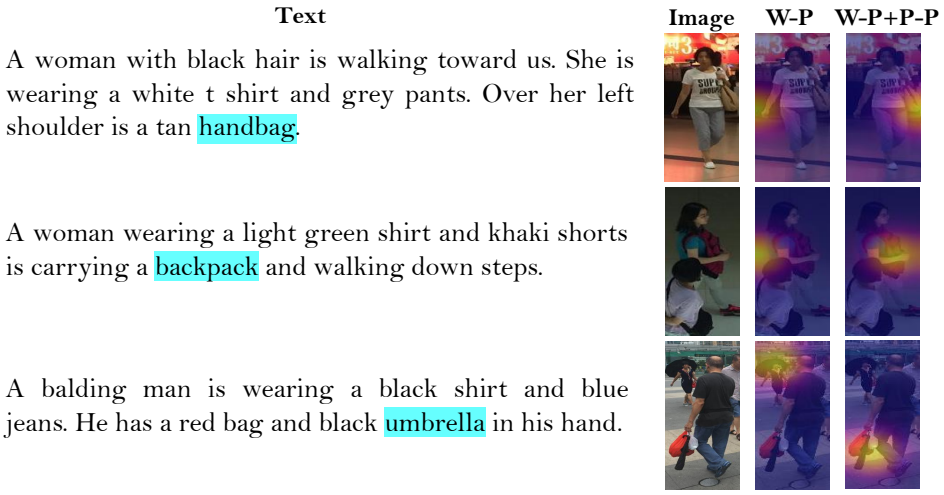
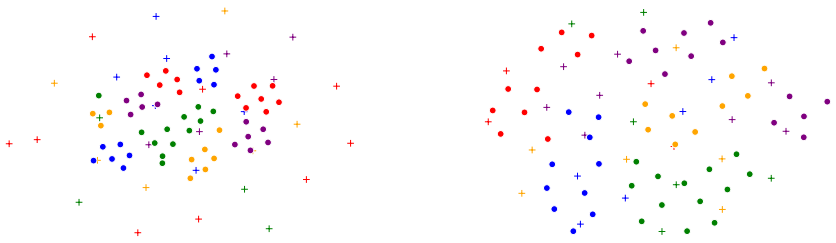


Figure 6.3: *Heatmaps of attention regions corresponding to the highlighted words from various components of HAITA-Net. Best viewed in color*

different shapes refer to different modality features. We show features for two phases: baseline features and features by our HAITA (from left to right).

Baseline phase: We visualize the baseline features in Figure 6.4(a). It shows that most of the data points are dispersed and some are totally merged into samples of another identity. It also shows that the baseline features of the same modality cluster well. However, image and text features are far apart. It is clearly evident that the baseline network is not enough to deal with cross-modal text-image features.

HAITA phase: On the other hand, Figure 6.4(b) shows that features from different modalities appear closer. For example, samples in red, green and blue color are closer as compared to baseline. This visualization shows the superiority of the proposed HAITA model over baseline. Also , it is clearly evident that the proposed network has strong ability to build an attentive relation between text-image to bridge the modality gap and make them closer as compared to the baseline phase.



(a) Baseline (b) HAITA

Figure 6.4: *2D visualization of the features. Different color stands for different class. The shape ★ denotes the feature extracted from the image, while the shape ● denotes the feature extracted from the text. Best viewed in color*

Visualization of subjective retrieval results: We show person search results with natural language descriptions in Figure 6.5. These results show that each retrieved image has regions that match corresponding parts of the text description. These cases are shown in the green bounding box. The erroneous results are marked in red bounding boxes.

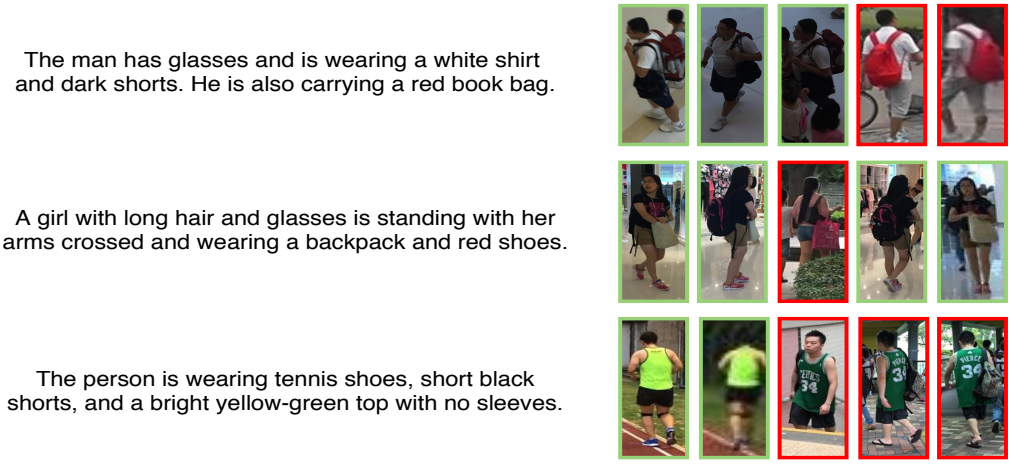


Figure 6.5: *Retrieval results. For each row, we show one text query and top-5 similar images. Best viewed in color*

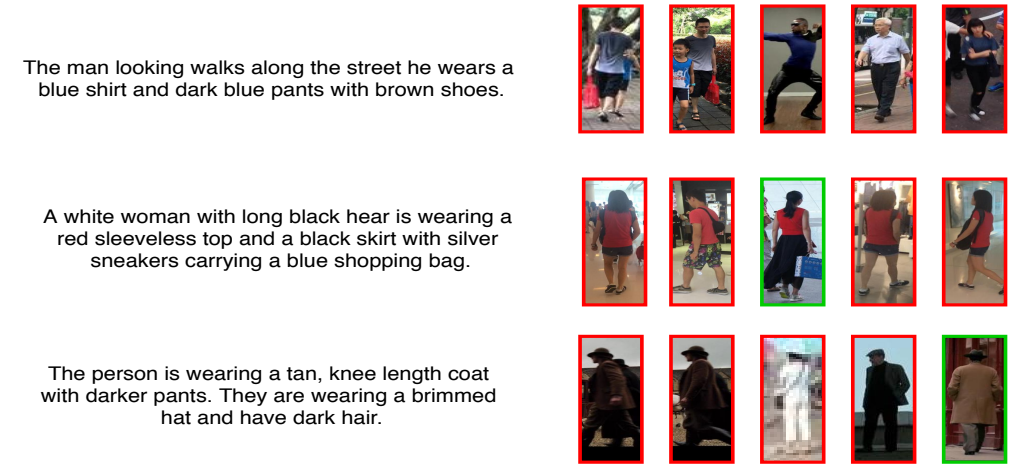


Figure 6.6: *Failure cases of retrieval results. Best viewed in color*

We also show failure retrieval cases in Figure 6.6. The failure results also retrieve images that are similar to the language descriptions. For example, in the first case, all of the retrieval images have a “blue shirt”. Similarly, in the second case, “red top” can be seen, and in the third example, and the last example contains "hat". The incorrect matches are highlighted in red color boxes.

6.3.7 Generalization Capability

To give more insight on HAITA-Net model, we not only address the problem of description Re-ID, but also demonstrate application of the proposed method to other object categories. We use Flickr-30K dataset for this purpose. To validate the superiority of the proposed method, we compare it with existing algorithms: VQA-A [174], DSPE [175], m-CNN [176], sm-LSTM [128], RRF-Net [124], CMPM+CMPC [118], NAR [177], DAN [178], A-GANet [121] and observe a better performance. We show the comparison results in Table 6.4 for text-to-image matching. We also show the comparison results with the state-of-the-art methods for image-to-text matching in Table 6.5. The significant improvement over the accuracy shows the generalization capability of the proposed method.

Table 6.4: *Comparisons on the Flickr30K dataset (Text-to-Image Re-ID). CMC-1, CMC-5, and CMC-10 (%) are reported.*

Method			**Flickr-30K**		
		Ref	CMC-1	CMC-5	CMC-10
Embedding	VQA-A	ECCV'16	24.9	52.6	64.8
	DSPE	CVPR'16	29.7	60.1	72.1
	m-CNN	ICCV'17	26.2	56.3	69.6
	sm-LSTM	CVPR'17	30.2	60.4	72.3
	RRF-Net	ICCV'17	35.4	68.3	79.9
	CMPM+CMPC	ECCV'18	37.3	65.7	75.5
	NAR	ICASSP'19	39.4	68.8	79.9
	Dual-Path [171]	ACM-TOMM'20	39.1	69.2	**80.9**
Attention	DAN	CVPR'17	39.4	69.2	79.1
	A-GANet	ACM MM'19	39.52	69.9	**80.9**
	HAITA-Net	**Ours**	**40.1**	**70.2**	79.9

Table 6.5: *Comparisons on the Flickr30K dataset (Image-to-Text Re-ID). CMC-1, CMC-5, and CMC-10 (%) are reported.*

Method	**Flickr-30K**		
	CMC-1	CMC-5	CMC-10
VQA-A	33.9	62.5	74.5
DSPE	40.3	68.9	79.9
m-CNN	33.6	64.1	74.9
sm-LSTM	42.5	71.9	81.5
CMPM+CMPC	40.3	66.9	76.7
HAITA-Net	**45.4**	**74.5**	**82.3**

This chapter discussed the contributions towards the emerging practical scenario of text-image heterogeneous Re-ID. In the next chapter, we conclude the research objectives with future directions.

Chapter 7

Conclusion and Future Works

7.1 Conclusion

This thesis focused on homogeneous and heterogeneous based person Re-ID. The main contributions of the thesis are:

• For homogeneous video Re-ID, we propose a novel end-to-end joint learning network. Our proposed CARF-Net uses a two stream triplet network which incorporates CNN, discriminative visual attention, RNN and multiple pooling layers. In order to preserve information from discriminative frames and to learn robust features, we apply multiple spatial, temporal and spatio-temporal fusion operations. Further, we use a multi loss method as an objective function to train the network. We rigorously evaluate the performance of multiple variants of the proposed network and conclude that CARF-Net gives best results. This is because the network exploits most informative regions for person re-identification through visual attention on frames and optical flow vectors, whereas, temporal and spatio-temporal pooling extracts most discriminative frames. The extensive comparison against several popular algorithms also show the efficiency of the proposed network. In addition, we empirically show that the proposed network is more robust to adversarial examples as compared to other algorithms. We also introduce a large novel aerial video dataset with several challenging situations.

• The second contribution deals with homogeneous image based Re-ID. We propose a novel deep re-id model based on an encoder-decoder network architecture for image based person re-id. Our hypothesis is that by augmenting the network with a multi-resolution decoder, the learned representation becomes richer and thus, generalizes better against severe challenges like pose variation, occlusion, illumination variation, and low resolution. Our elaborate quantitative and qualitative experiments prove this hypothesis. In addition, we propose a novel median based hybrid sampling strategy for mining of triplets. We observe that our sampling strategy achieves better clustering for positive and negative samples. We also propose test set augmentation technique with the reconstructed samples. The reconstructed samples are void of various challenges while preserving the identity specific information which indicates that the features possess both high discriminative and generative power for person re-id. Further, we validate the robustness against pose variation, partial occlusion as well as low resolution examples. We rigorously evaluate the performance of multiple variants of the proposed network and conclude that HDRNet achieves state-of-the-art performance.

• The third contribution presented in the thesis is the heterogeneous Re-ID. In this work, we propose a novel architecture for RGB-IR Re-ID with the goal of spectrum disentangled

representation learning. The architecture includes three major components. First, we introduce a disentanglement loss to segregate the identity and spectrum related features. Second, we design spectrum dispelling branch to learn identity cues and spectrum distilling branch to learn spectrum related information. Third, we apply an identity-dispeller over spectrum distilling branch to fool the identity classifier. The extensive experiments conducted on challenging RGB-IR datasets like SYSU-MM01 [5] and RegDB [6] show that the proposed network is robust enough to learn spectrum disentangled representation to perform RGB-IR Re-ID and outperforms state-of-the-art methods.

• In our final contribution, we not only address the problem of text-image Re-ID, but also demonstrate application of the proposed method to other object categories. We propose an end-to-end Hierarchical Attention Alignment Network (HA2-Net) for the text-image matching. This model is based on the hierarchical relations between the contexts of image and text. We apply an efficient threshold TF-IDF (T2FIDF) for the text to find out the most salient tokens. Furthermore, we design a multi-loss function that combines hierarchical weighted attention loss and cross-modal loss, and optimize our model in an end-to-end manner. Extensive experiments and analysis demonstrate the superiority of HA2-Net for efficiently learning discriminative image-text embeddings. The competitive performances on the CUHK-PEDES [10] and Flickr-30K [11] dataset outperform most previous methods significantly.

7.2 Future Works

The objective of this research is to develop robust algorithms for Re-ID under multi-camera networks for single and multi-modal data. Also, with the emerging research area of text-image matching for person category, several challenging problems are highlighted which require further investigation. We also believe that heterogeneous Re-ID will continue to be an active and promising research area with broad potential surveillance and other related applications. The future research directions are discussed in the following sections.

7.2.1 Video Re-ID through Autoencoders

Inspired from homogeneous image based HDRNet in chapter 4, we plan to mimic the similar behaviour in video network to reduce the complexity as well as to make learned features more generalizable. To achieve this goal, we plan to work on autoencoder network, comprising of an encoder and a sequential decoder for video based Re-ID. In this network, the encoder can be taken as the combination of CNN and RNN to exploit appearance and temporal feature vector. The output feature is then used as input to a decoder to reconstruct the features such that it preserves spatial location information of all the input frames. Additionaly, the goal of the decoder is to help the encoder to learn robust feature by introducing a generative task to the embedding layer so that it can become more generalizable to the unknown test data. Overall, the network can be optimized using multi-loss function– triplet verification and mean squared error loss.

7.2.2 Homogeneous Re-ID via Unsupervised Domain Adaptation

Unsupervised Re-ID has witnessed a surge in the number of works in the past few years. This is due to the fact that such system is easily scalable in a real world scenario unlike supervised system. However, unsupervised Re-ID also poses daunting challenges in terms of domain gap between a labeled source domain and unlabeled target domain. Though there has been a tremendous progress in unsupervised domain adaptation under classification setting, Re-ID has additional major challenge of completely unseen target labels [179–182]. This requires investigation of techniques which not only bridge domain gap but also address the challenge of totally unseen labels.

• To accomplish the goal of unsupervised Re-ID, source domain can be trained in a similar way as in HDRNet model. For target domain, the first step is to incorporate attributes-based Re-ID with the current setting of RGB-RGB Re-ID to parse the human semantically. Semantic attributes can be used as pseudo labels on target domain to address the challenges associated with source-target domain discrepancy.

• In order to learn the discriminative features and generalize the Re-ID model to the unsupervised target domain, we plan to exploit the disentanglement reconstruction model on source domain where learned representations can be disentangled into identity and non-identity related features. Further, the disentangled features can undergo reconstruction using a decoding layer to increase the generalization capability of the features. Then, the learned knowledge can be transferred to target domain (without label information) via fine-tuning. To learn the camera invariance property, training images of target domain can be style-transferred to each camera [16, 17, 183]. Further, original and cam-style generated images form the augmented training set and can be fed to target representation model. Once the domain-invariant feature is learned via fine-tuned model, we can perform cross-domain Re-ID by matching the query image and gallery images. A typical framework of the unsupervised model setting is shown in Fig. 7.1.

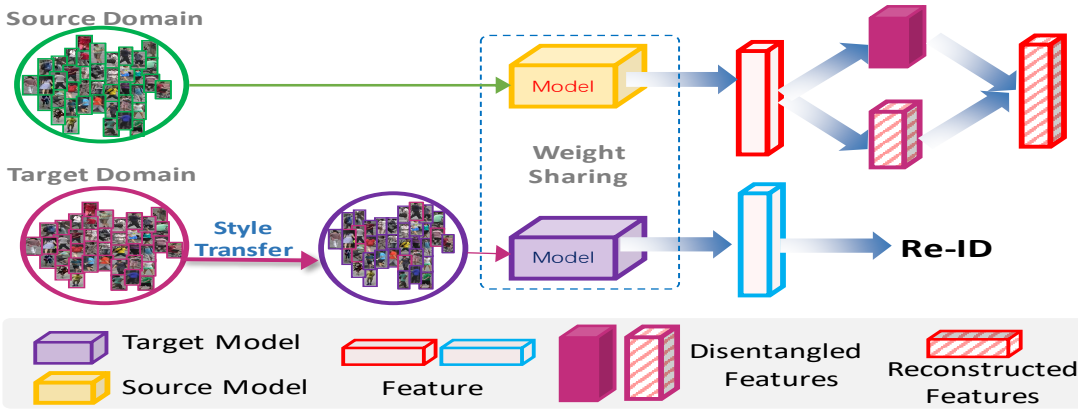


Figure 7.1: *UDA Model for Unsupervised Re-ID*

7.2.3 Other directions in Heterogeneous Re-ID

Due to the advancement of heterogeneous Re-ID in practical scenarios, there is high interest in this research area. However, many critical issues are still an open problem.

• The major success of homogeneous based Re-ID systems is due to the availability of large labeled datasets. However, there is a lack of large heterogeneous dataset with sufficient variations. The first possible future direction is to construct dataset with large number of samples to address the inter-modal discrepancy.

• There are many heterogeneous based applications such as Text-Image [10, 120, 127, 167] and Sketch-Image [184] Re-ID where human intelligence can be involved into the process of Re-ID. However, crowd-sourcing cues may be diverse to each other due to different views of each individual. For example, the descriptions of the text-annotations and sketch representations can vary a lot due to different persons' inputs. Hence, mining and integrating the rich information to search the query in the surveillance system remains a challenging problem. Here, the second possible future direction is towards the design of an automatic strategy [185, 186] to capture the rich contextual dependencies between the fragments of image and text/sketch in Re-ID models.

7.2.4 Usage of synthetic data

Leveraging the use of synthetic data is useful for cross-dataset person Re-ID. The lack of large-scale and diverse source training data limits the generalization ability of the learned models to unknown target domain. One possible solution is to automatically synthesize a large-scale Re-ID dataset similar to real surveillance but with virtual environments, and then use the synthesized person images to train a generalizable Re-ID model. Another way to reduce the domain gap is to fuse synthetic and real-world datasets to train the source models and then transfer the knowledge on the target domain. Few efforts [187–189] have been designed towards the creation of synthetic data for Re-ID domain. SOMAset [187] is a synthetic dataset with 50 3D-person models and 11 types of outfits. Bak *et al.*[188] also introduce SyRI dataset including 100 characters. This dataset is featured by rich lighting conditions. PersonX [189] dataset contains 1,266 3D characters and 273,456 bounding boxes taken with six cameras. However, their images only provide one character at a time under a single camera setting.

7.2.5 Regulation in person detection and processing

The performance of the Re-ID models heavily relies on successful person detection which is a challenging task. State-of-the-art [190–192] deep learning approaches (Faster R-CNN, R-FCN, SSD and YOLOv3) are used for the detection purpose. While the person detection has improved significantly over the recent years [193–195], crowd scenes remain particularly challenging for the detection and tracking tasks due to heavy occlusions, high person densities and significant variation in appearance. Recent works [196] suggest that there is significant further to increase object detection performance by utilizing even bigger datasets. These aspects can also be explored as a future direction of our work.

7.2.6 Advancement in Re-ID

Existing Re-ID models assume fixed gallery images and that a person appears in short continuous span of time. However, the real-world scenario is more challenging [197, 198] and there is a need to address long-term and open-world person Re-ID systems in which people probably appear after a certain amount of time and gallery images can be updated dynamically. Long-term and open-world person Re-ID are two of the main research directions.

Publications

Journal Articles

1. **Kansal, K.**, Subramanyam, A. V., Wang, Z., Satoh, S. I. (2020). SDL: Spectrum-Disentangled Representation Learning for Visible-Infrared Person Re-identification. IEEE Transactions on Circuits and Systems for Video Technology. DOI: 10.1109/TCSVT.2019.2963721. Github

2. **Kansal, K.**, Subramanyam, A. V., Prasad, D. K., Kankanhalli, M. (2019). CARF-Net: CNN attention and RNN fusion network for video-based person reidentification. Journal of Electronic Imaging, 28(2), 023036. DOI: https://doi.org/10.1117/1.JEI.28.2.023036. Github

3. **Kansal, K.**, Subramanyam, A. V. (2019). Hdrnet: Person re-identification using hybrid sampling in deep reconstruction network. IEEE Access, 7, 40856-40865. DOI: 10.1109/ACCESS.2019.2908344. Github

4. Saini, K., **Kansal, K.***, Subramanyam, A. V. (2019). Airborne visual tracking and reidentification system. Journal of Electronic Imaging, 28(2), 023003. DOI: https://doi.org/10.1117/1.JEI.28.2.023003. Github

   * - Corresponding Author

Conference Articles

1. **Kansal, K.**, Subramanyam, A. V., Wang, Z., Satoh, S. I. "Hierarchical Attention Alignment Network for Text-Image Matching", IEEE International Conference on Multimedia and Expo (ICME), 2021 (Submitted).

2. **Kansal, K.**, Subramanyam, A. V. (2019, September). Autoencoder Ensemble for Person Re-Identification. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 257-261). IEEE. DOI: 10.1109/BigMM.2019.00-15. Github

3. **Kansal, K.**, Subramanyam, A. V. (2018, October). Transfer learning of spatio-temporal information using 3D-CNN for person re-identification. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 923-928). IEEE. DOI: 10.1109/SMC.2018.00164. Github

References

[1] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011.

[2] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[3] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014.

[4] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 1092–1099, 2018.

[5] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5380–5389, 2017.

[6] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[7] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[8] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[9] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 3, 2017.

[10] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[11] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[12] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):523–536, 2018.

[13] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, volume 1, page 2, 2018.

[14] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. doi: 10.1109/TCSVT.2019.2898940. URL `http://dx.doi.org/10.1109/TCSVT.2019.2898940`.

[15] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019.

[16] 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.

[17] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28 (3):1176–1190, 2018.

[18] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, pages 407–419, 2019.

[19] Ju Dai, Pingping Zhang, Huchuan Lu, and Hongyu Wang. Video person re-identification by temporal residual learning. *arXiv preprint arXiv:1802.07918*, 2018.

[20] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, and Hui Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, pages 3552–3559, 2016.

[21] N McLaughlin, J Martinez del Rincon, and P Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[22] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[23] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng. Zero-shot person re-identification via cross-view consistency. *IEEE Transactions on Multimedia*, 18(2):260–272, 2016.

[24] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics*, 48(10):3006–3020, 2018.

[25] Kajal Kansal and A.V Subramanyam. Transfer learning of spatio-temporal information using 3d-cnn for person re-identification. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, 2018.

[26] Zheng Wang, Junjun Jiang, Yi Yu, and Shin'ichi Satoh. Incremental re-identification by cross-direction and cross-ranking adaption. *IEEE Transactions on Multimedia*, 21(9):2376–2386, 2019.

[27] Zheng Wang, Junjun Jiang, Yang Wu, Mang Ye, Xiang Bai, and Shin'ichi Satoh. Learning sparse and identity-preserved hidden attributes for person re-identification. *IEEE Transactions on Image Processing*, 2019. doi: 10.1109/TIP.2019.2946975. URL `http://dx.doi.org/10.1109/TIP.2019.2946975`.

[28] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018.

[29] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7501–7508, 2018.

[30] Kajal Kansal, AV Subramanyam, Zheng Wang, and Shin'ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE TCSVT*, 2020.

[31] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. Springer, 2014.

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[33] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[34] Kajal Kansal, Subramanyam Venkata, Dilip K Prasad, and Mohan Kankanhalli. Carf-net: Cnn attention and rnn fusion network for video-based person reidentification. *Journal of Electronic Imaging*, 28(2):023036, 2019.

[35] Kunal Saini, Kajal Kansal, and A Subramanyam Venkata. Airborne visual tracking and reidentification system. *Journal of Electronic Imaging*, 28(2):023003, 2019.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pages 346–361, 2014.

[37] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, 2017.

[38] Kajal Kansal and AV Subramanyam. Autoencoder ensemble for person re-identification. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 257–261. IEEE, 2019.

[39] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[40] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, pages 135–153, 2016.

[41] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01595*, 2016.

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[43] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. *ICCV*, 2017.

[44] Arne Schumann and Tobias Schuchert. Person re-identification in uav videos using relevance feedback. In *SPIE/IS&T Electronic Imaging*, pages 94070Z–94070Z, 2015.

[45] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Investigating open-world person re-identification using a drone. In *ECCV*, pages 225–240, 2014.

[46] Arne Schumann and Tobias Schuchert. Deep person re-identification in aerial images. In *SPIE/ Optics and Photonics for Counter terrorism, Crime Fighting, and Defence.*, page 99950M, 2016.

[47] Arne Schumann and Jurgen Metzler. Adapted deep feature fusion for person re-identification in aerial images. In *SPIE/ Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, page 106430L, 2018.

[48] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, pages 357–360, 2007.

[49] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. *arXiv preprint arXiv:1604.08683*, 2016.

[50] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 275–1, 2008.

[51] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015.

[52] Damien Simonnet, Michal Lewandowski, Sergio Velastin, James Orwell, and Esin Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *ECCV Workshop*, pages 423–432, 2012.

[53] Mark S Nixon and John N Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94 (11):2013–2024, 2006.

[54] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 75:77–89, 2018.

[55] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACMMM*, pages 420–428, 2017.

[56] Shiwei Zhang, Changxin Gao, Feifei Chen, Sihui Luo, and Nong Sang. Group sparse-based mid-level representation for action recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):660–672, 2016.

[57] Shiwei Zhang, Changxin Gao, Jing Zhang, Feifei Chen, and Nong Sang. Discriminative part selection for human action recognition. *IEEE Transactions on Multimedia*, 20(4):769–780, 2017.

[58] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[59] Khurram Soomro, Amir Roshan Zamir, and M Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[61] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[63] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016.

[64] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 562–572, 2019.

[65] 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 *CVPR*, pages 6776–6785, 2017.

[66] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661, 2017.

[67] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, pages 1983–1991, 2017.

[68] Zeng Yu, Tianrui Li, Ning Yu, Xun Gong, Ke Chen, and Yi Pan. Three-stream convolutional networks for video-based person re-identification. *arXiv preprint arXiv:1712.01652*, 2017.

[69] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017.

[70] Jianing Li, Shiliang Zhang, Jingdong Wang, Wen Gao, and Qi Tian. Lvreid: Person re-identification with long sequence videos. *arXiv preprint arXiv:1712.07286*, 2017.

[71] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016.

[72] M Feroz T Ali and Subhasis Chaudhuri. Maximum margin metric learning over discriminative nullspace for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 122–138, 2018.

[73] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

[74] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793, 2012.

[75] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[76] Ziyan Wu, Yang Li, and Richard J Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1095–1108, 2014.

[77] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[78] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[79] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016.

[80] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *arXiv preprint arXiv:1807.00537*, 2018.

[81] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *CVPR*, pages 5098–5107, 2018.

[82] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV)*, pages 365–381, 2018.

[83] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*, 2016.

[84] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[85] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, volume 2, 2017.

[86] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016.

[87] 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 *CVPR*, pages 1077–1085, 2017.

[88] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. *ECCV*, 2018.

[89] 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. *arXiv preprint arXiv:1709.09930*, 2017.

[90] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. Ca 3 net: Contextual-attentional attribute-appearance network for person re-identification. In *ACMMM*, pages 737–75. ACM, 2018.

[91] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, volume 1, page 2, 2018.

[92] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018.

[93] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5735–5744, 2019.

[94] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. *ECCV*, 2018.

[95] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019.

[96] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *AAAI*, 2018.

[97] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, pages 508–526. Springer, 2018.

[98] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016.

[99] 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 *ACMMM*, pages 192–196, 2016.

[100] Xuelin Qian1 Yanwei Fu, Yu-Gang Jiang, and Tao Xiang4 Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. ICCV, 2017.

[101] Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2018.

[102] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *ICPR*, pages 34–39, 2014.

[103] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[104] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[105] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018.

[106] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016.

[107] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[108] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[109] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856, 2019.

[110] Yang Li, Quan Pan, Suhang Wang, Haiyun Peng, Tao Yang, and Erik Cambria. Disentangled variational auto-encoder for semi-supervised learning. *Information Sciences*, 482:73–85, 2019.

[111] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1623–1632, 2017.

[112] Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pages 5656–5665, 2018.

[113] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

[114] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[115] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *International Conference on Biometrics (ICB)*, pages 174–181, 2018.

[116] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *European Conference on Computer Vision*, pages 606–624, 2018.

[117] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 677–683, 2018.

[118] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.

[119] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language person search with mutually connected classification loss. In *ICASSP*, 2019.

[120] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *IJCAI*, 2019.

[121] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *ACM Multimedia*, 2019.

[122] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

[123] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[124] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.

[125] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pages 4170–4178, 2016.

[126] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[127] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.

[128] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.

[129] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.

[130] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.

[131] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[132] Lin Chen, Hua Yang, Ji Zhu, Qin Zhou, Shuang Wu, and Zhiyong Gao. Deep spatial-temporal fusion network for video-based person re-identification. In *CVPRW*, pages 63–70, 2017.

[133] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[134] Aleksei Grigorev, Zhihong Tian, Seungmin Rho, Jianxin Xiong, Shaohui Liu, and Feng Jiang. Deep person re-identification in uav images. *EURASIP Journal on Advances in Signal Processing*, 2019(1):54, 2019.

[135] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, pages 4091–4099, 2015.

[136] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, pages 1–12, 2015.

[137] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, pages 4516–4524, 2015.

[138] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[139] Chenglin Liu, Tianlong Bao, and Ming Zhu. Part-based feature extraction for person re-identification. In *ICMLC*, pages 172–177, 2018.

[140] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov): 2579–2605, 2008.

[141] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[143] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[144] Raymond H Chan, Chung-Wa Ho, and Mila Nikolova. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on image processing*, 14(10):1479–1485, 2005.

[145] 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.

[146] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[147] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[148] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Local convolutional neural networks for person re-identification. In *ACMMM*, pages 1074–1082. ACM, 2018.

[149] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. *ICCV*, 2017.

[150] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. *ECCV*, 2018.

[151] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017.

[152] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

[153] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1089–1102, 2016.

[154] Jin Kyu Kang, Toan Minh Hoang, and Kang Ryoung Park. Person re-identification between visible and thermal camera images based on deep residual cnn using single input. *IEEE Access*, 7:57972–57984, 2019.

[155] Mang Ye. Cross-modal-re-id-baseline, 2018. URL `https://github.com/mangye16/Cross-Modal-Re-ID-baseline`.

[156] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *ACM Multimedia*, 2017.

[157] Xiang Xu, Ha A Le, Pengfei Dou, Yuhang Wu, and Ioannis A Kakadiaris. Evaluation of a 3d-aided pose invariant 2d face recognition system. In *IJCB*, 2017.

[158] J Ramos. Using tf-idf to determine word relevance in document queries. In *ICML*, 2003.

[159] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[160] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[161] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2), 2016.

[162] 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 1–7. Citeseer, 2007.

[163] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian conference on computer vision*, pages 31–44. Springer, 2012.

[164] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[165] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[166] Zhong Ji, Shengjia Li, and Yanwei Pang. Fusion-attention network for person search with free-form natural language. *PRL*, 2018.

[167] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, pages 56–73. Springer, 2018.

[168] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020.

[169] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[170] Ammarah Farooq, Muhammad Awais, Fei Yan, Josef Kittler, Ali Akbari, and Syed Safwan Khalid. A convolutional baseline for person re-identification using vision and language descriptions. *arXiv preprint arXiv:2003.00808*, 2020.

[171] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[172] Vikram Shree, Wei-Lun Chao, and Mark Campbell. Interactive natural language-based person search. *IEEE Robotics and Automation Letters*, 5(2):1851–1858, 2020.

[173] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. In *AAAI*, pages 11189–11196, 2020.

[174] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.

[175] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[176] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.

[177] Zang W Wang B Liu C, Mao Z. A neighbor-aware approach for image-text matching. In *ICASSP*, 2019.

[178] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.

[179] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018.

[180] Yixiao Ge, Feng Zhu, Rui Zhao, and Hongsheng Li. Structured domain adaptation for unsupervised person re-identification. *arXiv preprint arXiv:2003.06650*, 2020.

[181] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–19, 2020.

[182] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*, 2018.

[183] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[184] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 609–617, 2018.

[185] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[186] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096, 2019.

[187] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 167:50–62, 2018.

[188] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. *arXiv preprint arXiv:1804.10094*, 2018.

[189] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 608–617, 2019.

[190] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[191] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[192] Rayson Laroca, Evair Severo, Luiz A Zanlorensi, Luiz S Oliveira, Gabriel Resende Gonçalves, William Robson Schwartz, and David Menotti. A robust real-time automatic license plate recognition based on the yolo detector. In *2018 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2018.

[193] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[194] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, 2009.

[195] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[196] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

[197] Srikrishna Karanam, Eric Lam, and Richard J Radke. Rank persistence: Assessing the temporal performance of real-world person re-identification. In *Proceedings of the 11th International Conference on Distributed Smart Cameras*, pages 157–162, 2017.

[198] Meng Zheng, Srikrishna Karanam, and Richard J Radke. Measuring the temporal behavior of real-world person re-identification. *arXiv preprint arXiv:1808.05499*, 2018.