Spring 5-22-2021

Person Re-identification And An Adversarial Attack And Defense For Person Re-identification Networks

Yu Zheng
Syracuse University, yzheng04@syr.edu

ABSTRACT

PERSON re-identification (ReID) is the task of retrieving the same person, across different camera views or on the same camera view captured at a different time, given a query person of interest. There has been great interest and significant progress in person ReID, which is important for security and wide-area surveillance applications as well as human computer interaction systems. In order to continuously track targets across multiple cameras with disjoint views, it is essential to re-identify the same target across different cameras.

This is a challenging task due to several reasons including changes in illumination and target appearance, and variations in camera viewpoint and camera intrinsic parameters. Brightness transfer function (BTF) was introduced for inter-camera color calibration, and to improve the performance of person ReID approaches. In this dissertation, we first present a new method to better model the appearance variation across disjoint camera views. We propose building a codebook of BTFs, which is composed of the most representative BTFs for a camera pair. We also propose an ordering and trimming criteria, based on the occurrence percentage of codeword triplets, to avoid using all combinations of codewords exhaustively for all color channels, and improve computational efficiency. Then, different from most existing work, we focus on a crowd-sourcing scenario to find and follow person(s) of interest in the collected images/videos. We propose a novel approach combining R-CNN based person detection with the GPU implementation of color histogram and SURF-based re-identification. Moreover, GeoTags are extracted from the EXIF data of videos captured by smart phones, and are displayed on a map together with the timestamps.

With the recent advances in deep neural networks (DNN), the state-of-the-art performance of person ReID has been improved significantly. However, latest works in adversarial machine learning have shown the vulnerabilities of DNNs against adversarial examples, which are carefully crafted images that are similar to original/benign images, but can deceive the neural network models. Neural network-based ReID approaches inherit the vulnerabilities of DNNs. We present an effective and generalizable attack model that generates adversarial images of people, and results in very significant drop in the performance of the existing state-of-the-art person re-identification models. The results demonstrate the extreme vulnerability of the existing models to adversarial examples, and draw attention to the potential security risks that might arise due to this in video surveillance. Our proposed attack is developed by decreasing the dispersion of the internal feature map of a neural network. We compare our proposed attack with other state-of-the-art attack models on different person re-identification approaches, and by using four different commonly used benchmark datasets. The experimental results show that our proposed attack outperforms the state-of-art attack models on the best performing person re-identification approaches by a large margin, and results in the most drop in the mean average precision values.

We then propose a new method to effectively detect adversarial examples presented to a person ReID network. The proposed method utilizes parts-based feature squeezing to detect the adversarial examples. We apply two types of squeezing to segmented body parts to better detect adversarial examples. We perform extensive experiments over three major datasets with different attacks, and compare the detection performance of the proposed body part-based approach with a ReID method that is not parts-based. Experimental results show that the proposed method can effectively detect the adversarial examples, and has the potential to avoid significant decreases in person ReID performance caused by adversarial examples.

PERSON RE-IDENTIFICATION AND AN ADVERSARIAL ATTACK AND DEFENSE FOR PERSON RE-IDENTIFICATION NETWORKS

By

Yu Zheng
B.S., Shanghai University, Shanghai, China, 2008
M.S., Syracuse University, Syracuse, NY, USA, 2012

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Syracuse University
May 2021

For all the loving support, motivation, and reasoning:
this dissertation is dedicated to my wonderful wife Ming Li

Acknowledgements

Through the process of writing my dissertation, I feel like the Ph.D. journey is very much like a roller coaster. It has peaks and valleys, but all we need to do is keeping head up and accumulating the energy for the future. Over the years, there are joys, tears, excitements, frustration...I still remember how shocking it was when our first paper got rejected. Great frustration and self-doubts drag me into the mud of sadness. I still remember how exciting it was when I was presenting my paper in front of more than 100 experts in the conference. I went through it over and over on the night before to make sure I can talk smoothly during the presentation. There are many imperishable memories like those during my whole journey, and I know that I would not be able to achieve anything without the help of many people that I would like to take this opportunity to recognize here.

First and foremost, I would like to express my deepest gratitude to my advisor Dr. Senem Velipaşalar, for her valuable advice, sleepless nights working over conference and journal submissions, and the guidance on this very challenging journey. Once we had a deadline at 5am, she helped us with all editing and submission, I went back to lab at 10am and saw her finishing teaching for an early class already. Moreover, she was pregnant at that time with her daughter. It is the first time in my life that I have seen such dedication for students and the enthusiasm in research. The spark of inspiration will fuel me for the rest of my life.

I am indebted to Dr. Jianshun Zhang for serving as the oral examination chair for my defense and Drs. Asif Salekin, Sara Eftekharnejad, Jay K. Lee, and Garret Katz for being a part of my defense committee. Their input and suggestions helped evolve this thesis into what it is now. Also, I would like to thank Dr. Reza Zafarani for being in the proposal defense committee, it is a great pity to miss you for my defense.

Special thanks are extended to the members of the Smart Vision Systems Lab-

oratory at Syracuse University, especially Mauricio Casares, Anvith Mahabalagiri, Akhan Almagambetov, Koray Özcan, Maria Cornacchia, Chuang Ye, Burak Kakillioğlu, Danushuka Bandara, Yantao Lu, Jiyang Wang, Fatih Altay, Weiheng Chai, Jiajing Chen, Feng Wang, and Chenbin Pan for their help and insightful discussions.

I am grateful for Dr. Jae Oh, Ms. Cynthia V Salanger for the help in granting my assistantship, program status and other paperwork. Also, I would like to thank the National Science Foundation (NSF), and the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, for supporting this research in part (under grants CNS-1739748, CNS-1816732, DE-AR0000940).

I could not have finished (or started) this Ph.D. without the inspiration and motivation from my grandmother Ruihua Li, grandfather Songzheng Zheng, mother Yiqin Zheng and father Guiyao Cui. The sacrifices they made provided me with a happy childhood and college years. I am forever indebted to them for all that they have done for me.

Finally, I would like give my sincerest gratification to my wife Ming Li for her unconditional love and support that she always give me. It has been a long and winding road filled with roadblocks, but we are gratefully at the end of it. Thank you, Ming, for being with me over the years of up and down, joys and tears. I dedicate this dissertation to you.

■

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Given an image/video of a person of interest, person re-identification (ReID) is the process of identifying same person in the images/video captured by a different camera. Re-identification is indispensable in establishing the consistent labeling across multiple cameras with disjoint views or even within the same camera to connect broken trajectories or re-establish lost tracks. Person ReID is a difficult problem due to several reasons, including large variations in illumination and target appearance, and variations in camera intrinsic parameters and viewpoint. Person re-identification across different camera views has two main parts; (1) detecting the people in the scene and (2) re-identifying them in other camera views. Various person ReID approaches have been proposed in the past [3], which can be classified into different categories. There have been methods based on distance learning [4, 5, 6, 7], on feature design and selection [8, 9, 10], and on mid-level feature learning [11, 12, 13, 14].

Porikli [15] proposed the Brightness Transfer Function (BTF) for inter-camera color calibration. Later, Javed et al. [16] proposed Mean Brightness Transfer Function (mBTF) by projecting BTF to low subspace and computing the average appearance similarity to improve the performance. Prosser et al. [17] proposed the Cumulative Brightness Transfer Function (cBTF) to compensate the illumination change over time. Datta et al. [2] presented the Weighted-BTF (wBTF), which weighs the BTFs for K training images, whose background areas are close to the background of a test image. Bhuiyan et al. [18] presented the Minimum Multiple Brightness Transfer Function (Min-MCBTF) to model the appearance variation by using a learning approach. However, it is assumed that multiple consecutive images are available for training, which is not the case for the commonly used benchmark datasets.

In Chapter 2, we present an approach to obtain a better color calibration and brightness transfer across disjoint camera views, and in turn to improve the person ReID performance. Any person ReID approach incorporating color or brightness histograms can benefit from the proposed method, since it improves the brightness transfer. We propose to build a codebook of BTFs for a camera pair. The codebook contains the most representative BTFs codewords for each color channel. We propose to use the Chain Code Histogram (CCH) for measuring the similarity between two BTFs. Moreover, to improve the performance even further, we present a different approach to segment a person from the background in dataset images. In addition, to avoid using all combinations of codewords exhaustively for all color channels, and thus improve the computational efficiency, we present an ordering and trimming criteria, which is based on the occurrence percentage of codeword triplets.

Large camera networks are increasingly being deployed in public places such as airports, subways, campuses and office buildings. These cameras are usually installed across wider areas with non-overlapping fields of view (FOV) to provider better coverage. However, with these setups, tracking of person(s) is still limited to the areas where the cameras are installed. Thanks to the widespread use of smart phones, crowdsourcing of videos has the potential to provide a much larger coverage for longer periods of time. For instance, a video or image captured by someone at a remote location (with no pre-installed camera) can have the person of interest in it. If it is possible to have access to large number of videos captured at different locations over time, this would potentially allow following a person over a much larger area for extended periods of time. However, the extremely large amount of data captured in these scenarios makes it impossible for humans to perform manual analysis, and necessitates autonomous analysis of data. Video analysis enable long term characterization and modeling of the people in the scene. Such application is required for high-level surveillance tasks and make them even smarter [19].

In Chapter 3, different from most of the existing work, we focus on a crowdsourcing scenario to find and follow person(s) of interest in the collected images/videos. We present an approach combining R-CNN based person detection with the GPU implementation of color histogram and SURF-based re-identification. Moreover, the GeoTags are extracted from the EXIF data of videos captured by smart phones, and these locations are displayed on a map together with the time-stamps for a spatiotemporal representation/visualization of the trajectory.

With the advancement of deep neural networks(DNN) and increasing demand for intelligent video surveillance, neural network-based methods have been applied to the person ReID problem across different camera views, achieving state-of-the-art performance. However, it has been shown that neural networks are vulnerable to adversarial examples, which are carefully designed to be close to the original input, but can easily deceive these networks. The ReID approaches employing neural network-based models inherit the vulnerabilities of DNNs. Adversarial examples [20, 21, 22] have been extensively investigated in image classification [22, 23], object detection [24, 25, 26] and semantic segmentation [27, 24] tasks. However, relatively less attention has been paid to the robustness of person ReID models. In Chapter 4, we present an attack model that generates adversarial images of people, and demonstrate the effectiveness of this model by attacking multiple state-of-the-art person ReID models. We also compare the performance of the presented attack approach with other state-of-the-art attack models via an extensive set of experiments on various person ReID benchmark datasets. One of our goals is to demonstrate the extreme vulnerability of multiple state-of-the-art person ReID approaches to this attack, and draw attention to the existing security risks. In addition, we use different network models (different from the model used by the victim ReID networks) as the source model to generate the adversarial examples and show the effectiveness and generalizability of our attack approach. We also analyze the effect of the perturbation budget on the attack performance. The experimental results show that our proposed attack outperforms the state-of-art attack models on the best performing person re-identification approaches by a large margin, and results in the most drop in the mean average precision values.

While adversarial attacks have shown the vulnerability of DNNs, various defense approaches against adversarial examples have been proposed [28, 29, 30, 31, 32]. However, defending ReID networks against adversarial attacks is still relatively unexplored. In Chapter 5, we propose a new method to effectively detect adversarial examples presented to a person ReID network by utilizing part-based feature squeezing. Feature squeezing was proposed for detecting the adversarial examples in image classification task with efficient computation compared to other iterative methods [32]. We show that by applying the feature squeezer on top of the body parts-based ReID model, the AE detection performance can be further improved compared to using a network that is not based on parts. We apply two types of squeezing to segmented body parts. Experimental results show that the proposed method can effectively detect the adversarial examples, and has the potential to avoid significant decreases in person ReID performance caused by adversarial examples. More specifically, we show that by detecting the adversarial examples, the mean average precision (mAP) of person ReID models can be increased compared to not detecting AEs at all. With the PCB model, the mAP after AE detection can reach close to 70% (compared to an mAP of 22.6 before AE detection).

Finally, in Chapter 6, we provide a conclusion together with future work directions. The research conducted during this Ph.D. study resulted in several publications, including respected Institute of Electrical and Electronics Engineers (IEEE) journals and international conference proceedings, listed below:

Publications

Peer-reviewed Journal Papers

- Y. Zheng, Y. Lu and S. Velipasalar, "An Effective Adversarial Attack on Person Re-Identification in Video Surveillance via Dispersion Reduction," IEEE Access, vol. 8, pp. 183891-183902, Sep. 2020.

- M Cornacchia, B Kakillioglu, Y Zheng, S Velipasalar, "Deep learning-based obstacle detection and classification with portable uncalibrated patterned light," IEEE Sensors Journal 18 (20), 8416-8425, 2018.

- M. Cornacchia, K. Ozcan, Y. Zheng and S. Velipasalar, "A Survey on Activity Detection and Classification Using Wearable Sensors," IEEE Sensors Journal, vol. 17, issue: 2, pp. 386-403, Jan. 2017.

Peer-reviewed Conference Papers

- Y. Zheng and S. Velipasalar, "Part-based Feature Squeezing to Detect Adversarial Examples In Person Re-Identification Networks", submitted, 2021.

- Y. Zheng, K. Ozcan and S. Velipasalar, "A Codebook of Brightness Transfer Functions for Improved Target Re-Identification across Non-Overlapping

Camera Views," Proc. of the IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 166–170, Nov. 2017.

- Y. Zheng, Zhenhua (Jimmy) Chen, S. Velipasalar and J. Tang, "Person Detection and Re-identification Across Multiple Images and Videos Obtained via Crowdsourcing," Proc. of the International Conference on Distributed Smart Cameras (ICDSC), pp. 178-–183, Sept. 2016.

- Y. Zheng, K. Ozcan, S. Velipasalar, H. Shen, Q. Qiu, "Energy Efficient Tracking by Dynamic Voltage and Frequency Scaling on Android Smart Phones," Proc. of the ACM Int'l Conf. on Distributed Smart Cameras (ICDSC), pp. 1–6, Nov. 2014.

- Y. Zheng, C. Ye, S. Velipasalar and M. C. Gursoy, "Energy Efficient Image Transmission using Wireless Embedded Smart Cameras," Proc. of the IEEE Int'l Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 62–67, Aug. 2014.

- Y. Zheng, A. Mahabalagiri and S. Velipasalar, "Detection of Moving People with Mobile Cameras by Fast Motion Segmentation," Proc. of the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), pp. 1–6, Oct. 29- Nov. 1, 2013.

- C. Ye, Y. Zheng, S. Velipasalar and M. C. Gursoy, "Energy-Aware and Robust Task (Re)Assignment in Embedded Smart Camera Networks," Proc. of the IEEE International Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 123–128, August 2013.

Chapter 2
Improving Person Re-identification across Disjoint Camera Views Using Codebook of Brightness Transfer Functions

2.1 Introduction

In order to continuously track targets across multiple cameras with disjoint views, it is essential to re-identify the same target on different camera views. However, this is a very challenging task due to several reasons, including changes in illumination and target appearance, and variations in camera intrinsic parameters and viewpoint.

Various person re-identification (ReID) approaches have been proposed in the past [3], which can be classified into different categories. There have been methods based on distance learning [33, 34, 35, 4, 5, 6, 7], on feature design and selection [36, 8, 37, 38, 9, 39, 10], and on mid-level feature learning [40, 41, 11, 12, 13].

Cheng and Piccardi [42] apply a cumulative color histogram transformation and also employ an incremental major color spectrum histogram representation while assuming that a color mapping function is known a priori. Trajectory matching is employed, and height estimation and illumination-tolerant color representation are used by Madden and Piccardi [43]. Chae and Jo [44] employ a Gaussian Mixture Model (GMM), and use a ratio of the GMMs to identify the same person.

Porikli [15] proposed the Brightness Transfer Function (BTF) for inter-camera color calibration. Later, Javed et al. [16] proposed Mean Brightness Transfer Function (mBTF) by projecting BTF to low subspace and computing the average appearance similarity to improve the performance. Prosser et al. [17] proposed the Cumulative Brightness Transfer Function (cBTF) to compensate the illumination change over time. Datta et al. [2] presented the Weighted-BTF (wBTF), which weighs the BTFs for K training images, whose background areas are close to the background of a test image. Bhuiyan et al. [18] presented the Minimum Multiple Brightness Transfer Function (Min-MCBTF) to model the appearance variation by using a learning approach. However, it is assumed that multiple consecutive images are available for training, which is not the case for the commonly used benchmark datasets.

More recently, Liao et al. [45] proposed Local Maximal Occurrence (LOMO) and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) for person ReID. Chen et al. [46] formulated a new view-specific person reID framework, referred to as camera correlation aware feature augmentation (CRAFT). In this framework, cross-view feature adaptation is performed by measuring cross-view correlation from visual data distribution and carrying out adaptive feature augmentation. Matsukawa et al. [47] proposed the GOG, which first divides an image into horizontal strips. Local patches in the strips are modeled using a Gaussian distribution. Köstinger et al. [48] propose an approach, referred to as KISSME, which is a statistical inference perspective to address the problem of metric learning. As will be mentioned below, we evaluate the performance improvement provided by our proposed approach when it is used in combination with all these four state-of-the-art approaches [45][46][47][48].

In this dissertation, we propose a novel approach to obtain a better color calibration and brightness transfer across disjoint camera views, and in turn to improve the person ReID performance. As will be shown by our experimental results, any person ReID approach incorporating color or brightness histograms can benefit from the proposed method, since it improves the brightness transfer. For instance, most of the aforementioned methods incorporate color or brightness histograms. Using our proposed method first for brightness transfer, and then incorporating other features and using a different distance metrics is expected to increase the accuracy even further. Our experimental results support this argument as will be discussed in detail below.

We propose to build a codebook of BTFs for a camera pair. The codebook contains the most representative BTFs -codewords- for each color channel. We propose to use the Chain Code Histogram (CCH) for measuring the similarity between two BTFs. Moreover, to improve the performance even further, we present a different approach to segment a person from the background in dataset images. In addition, to avoid using all combinations of codewords exhaustively for all color channels, and thus improve the computational efficiency, we present an ordering and trimming criteria, which is based on the occurrence percentage of codeword triplets.

In this dissertation, we apply our proposed approach in combination with four different approaches, namely CRAFT, LOMO+XQDA, GoG and WHOS+KISSME, on the person ReID problem, and evaluate its effectiveness. The ReID performance is compared for all methods on VIPeR, CUHK01 as well as CUHK03 datasets. In addition, a detailed pseudocode of the proposed algorithm is presented.

To emphasize, the goal here is not a careful selection of different image features, or combinations of patches, or distance metrics, but instead to show the effect of performing better color calibration and brightness transfer, on the ReID performance when the same features and distance metric are used. The proposed approach was incorporated into four different state-of-the-art person re-identification methods, including LOMO+XQDA and CRAFT, and an increased top rank matching rate has been obtained for all methods and on all datasets, supporting the argument that the proposed method provides an improved brightness transfer across different camera views, and any person re-identification approach incorporating color/brightness histograms can benefit from it. Moreover, the performance improvement provided by the proposed method becomes more pronounced when the training dataset size becomes smaller. Thus, the proposed method becomes more preferable when there is not enough training data.

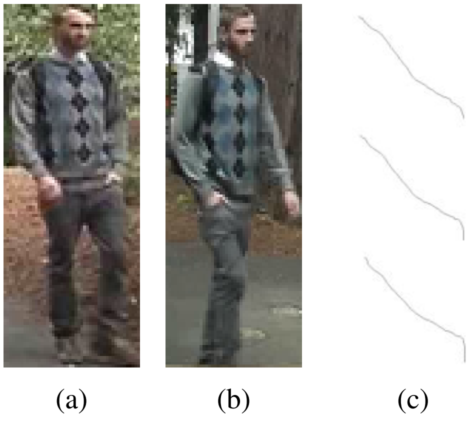
The rest of this chapter is organized as follows. The details of the proposed method are described in Section 3.2. Experimental results are presented in Section 5.4, and the chapter is concluded in Section 2.4.

2.2 Proposed Method

In order to model the variations in object appearances between different and non-overlapping camera views, we propose to build a codebook of BTFs from a training set. The BTFs are computed by using the segmented out foreground (target) regions. The details of how we perform this segmentation are described in Section 2.2.1.

The built codebook contains the most representative BTFs, between a camera pair, as codewords. The codewords are obtained for each of the red, green and blue color channels. Figure 2.1c shows three BTFs, which were extracted for Red (R), Green (G) and Blue (B) color channels of an example corresponding image pair (Fig. 2.1a and 2.1b) from the VIPeR dataset [1]. When we inspected the BTFs computed for different corresponding images from the same camera pair, we observed that while some BTFs looked similar, there were others that looked quite different as seen in Fig. 2.3. By constructing a codebook of BTFs, we are able to capture this variation among BTFs for the same camera pair, and obtain a model representing various illumination changes as targets move from one camera to another.

Depending on the color variation that is observed in the training set, each color channel might have different number of codewords. A new codeword is generated only if the newly computed BTF, between a training pair, is different enough from the existing BTFs in the codebook. If a newly computed BTF matches to one of the existing ones, the frequency of use of that codeword is incremented by one. Thus, along with a BTF (codeword), the frequency of use of that BTF during training is also saved in the codebook. The details of how each codeword is generated are described in Section 2.2.2.



(a) (b) (c)

Fig. 2.1: (a) and (b) An example corresponding image pair from the VIPeR dataset [1], and (c) the extracted BTFs for R, G, B color channels from top to bottom.

2.2.1 Segmenting Persons from Background

First, we employ Normalized Cuts [49] to over-segment an image similar to [2]. In [2], the foreground model is initialized by using segments that are in the center of the image. It then uses the minimum Euclidean distance between any segment and the center segments, and the Bhattacharyya distance between their color histograms to label the foreground regions. In this dissertation, we present a different approach for this step. Using the fact that the person is at the center of the image, in our approach, the segments are sorted by their minimum distance to the image corners. Then, segments are added one by one to the background region as long as the background is smaller than the 40% of the entire image. This way, the proposed approach can remove much more, if not all, of the background compared to the approach in [2], as shown in Fig. 2.2. In Fig. 2.2, columns (b) and (c) show the segmented foreground regions with the proposed method and the one in [2], respectively. Removing more of the background, especially when the background has very different colors compared to the target, is more important to obtain better transfer functions across a pair of camera views.

2.2.2 Codebook Construction

Brightness Transfer Functions (BTF) were proposed by Porikli [15] and have been used for inter-camera color calibration. We follow a similar approach in order to compute the BTF between an image pair (I_k^a, I_k^b), where a and b refer to two different cameras, and k is the index of the matched image pair in the dataset. We compute cumulative histograms of H_{k,c_i}^a and H_{k,c_i}^b for each color channel c_i. Then, the correlation matrix C is obtained between two histograms to represent the bin-wise mutual distances. The shortest cost path, connecting the top left vertex to the bottom right, is the original BTF defined in [15]. Figure 2.1c shows three BTFs,

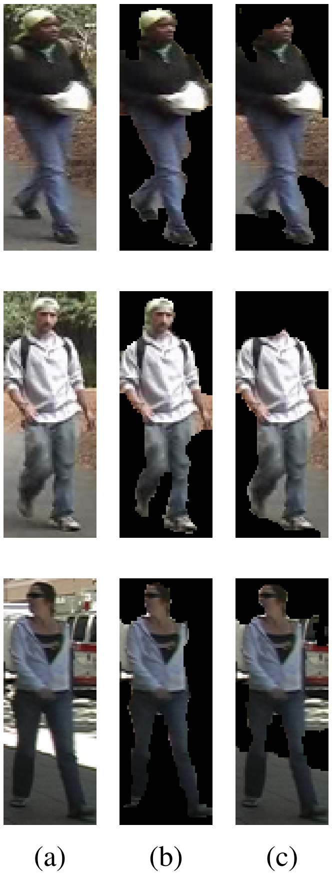


Fig. 2.2: Example outputs of foreground - background segmentation (a) Original image, (b) segmentation result with the proposed approach, (c) segmentation result with the approach in [2].

which were extracted for R, G and B color channels of a pair of corresponding images (Fig. 2.1a and 2.1b) from the VIPeR dataset [1]. Codewords are generated based on these BTFs extracted from the training set.

The steps of the algorithm for creating the codewords, for each color channel, are outlined in Alg 1. For the first corresponding image pair, i.e. when the codebook is empty, the BTFs computed for each channel are saved as the first codewords cw_1^R, cw_1^G, cw_1^B. Then, for every incoming corresponding image pair, the BTF is calculated, and compared to the existing codewords in the codebook. If the newly computed BTF is different enough from the existing ones, i.e. it does not match to any existing codeword, a new codeword is created. If not, i.e. when it matches to an existing codeword, the frequency of use $f_{cw_i}^R, f_{cw_i}^G or f_{cw_i}^B$ of that codeword is incremented by one.

We use chain code histograms (CCH) to measure the similarity between two BTFs. CCHs provide a scale and translation invariant shape descriptor for object contours in binary images [50]. It has also been shown that printed letters can be classified based on their CCHs [51]. Figures 2.1c and 2.3 show some example BTFs computed between image pairs. As can be seen from these figures, BTFs can be treated as binary images containing a curve, and CCHs can be used as a shape descriptor. Using the pixel locations on the curve, we use eight directions, starting from the bottom left corner, to compute the chain codes. After obtaining the chain codes for all pixel locations, we build an 8-bin histogram of the chain codes based on the frequency of occurrence for each direction. Two CCHs are compared by using the dissimilarity distance in Eq. (2.1),

$$D = 1 - \frac{\displaystyle\sum_{i=0}^{N-1}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\left[\displaystyle\sum_{i=0}^{N-1}(r_i - \bar{r})^2 \sum_{i=0}^{N-1}(s_i - \bar{s})^2\right]}}, \tag{2.1}$$

$$\bar{r} = \frac{1}{N}\sum_{i=0}^{N-1}(r_i),\ \ \bar{s} = \frac{1}{N}\sum_{i=0}^{N-1}(s_i),$$

where r and s are N-dimensional feature vectors. In this case, the CCH vector is of size 8. The dissimilarity distance in Eq. (2.1) is zero when two histograms are identical, and it increases as the histograms start to deviate. In our proposed method, a new codeword is created if the dissimilarity distance is greater than a threshold ρ. As mentioned in Section 5.4, the same value for ρ was used in all of our evaluations and experiments.

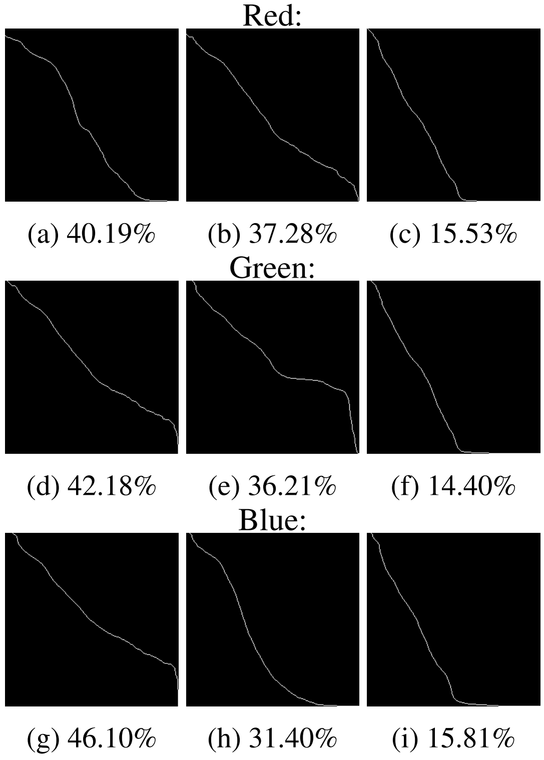
As outlined in Alg 1, while building the codebook, the frequency of codeword occurrence and frequency of occurrence of codeword triplets (for each channel) are

Algorithm 1 Construction of Codebook of BTFs

```
Initialize codebook CB=$\phi$
for $Each\ corresponding\ image\ pair$ do
    for $Each\ j = R, G, B$ do
        Compute BTF and its CCH
        if (CB = $\phi$) then
            Create first codeword $cw^j$ from BTF
            $f^j_{cw_1} = 1;$
        else
            match =0;
            for $Each\ codeword$ do
                compare CCH distance $D$
                if $D \leq \rho$ for $cw_i$ then
                    $f^j_{cw_i} + +$; match =1;
                end if
            end for
            if match is 0 then
                Create new codeword $cw^j_k$ from BTF
                $f^j_{cw_k} = 1;$
            end if
        end if
    end for
end for
```

also obtained by incrementing counters every time a newly computed BTF matches to an existing codeword. More specifically, if the BTFs of the each color channel of the next image pair are matched to the i^{th}, j^{th} and k^{th} codewords of the red, green, and blue channels, respectively, then the triplet (i,j,k) is incremented by one. This number is used to calculate a percentage of occurrence for different triplets.

Once the codebook is formed (during training), it is used during testing for performance analysis. One option during testing is to exhaustively use all combinations of codewords, for all channels, for image matching. However, to avoid this combinatorial approach and improve computational efficiency, we propose a trimming criteria, which is based on the occurrence percentage of codeword triplets. The codeword triplets are rank ordered in descending order of occurrence percentage, and only the top N triplets for which the sum of the percentages of occurrence



Red:

(a) 40.19% (b) 37.28% (c) 15.53%

Green:

(d) 42.18% (e) 36.21% (f) 14.40%

Blue:

(g) 46.10% (h) 31.40% (i) 15.81%

Fig. 2.3: Top three most frequently occurring BTFs for red, green and blue channels when $\rho = 0.05$.

reaches at least 75% are kept. This way, exhaustive combination of all possible codewords for each channel is avoided. As presented in Sect. 5.4, we also performed experiments by using all possible combinations of codewords for different channels, and observed that this does not improve the performance significantly.

Fig. 2.3, shows top three most frequent BTFs for each color channel together with their percentage of occurrence. Both x-axis and y-axis values are ranging from 0 to 255 for BTF representations.

2.3 Experimental Results

As mentioned above, our purpose here is not a careful selection of different image features, or combinations of patches, or distance metrics, but instead to show the effect of performing better color calibration and brightness transfer, on the ReID performance when the same features and distance metric are used. Any target ReID approach, which relies on appearance and color or brightness histograms, can benefit from the proposed approach, since it provides an improved brightness transfer. Using the proposed method first for color calibration, and then incorporating other features and using different distance metrics is expected to increase the accuracy even further. In order to support this argument, we have performed extensive experiments with four different state-of-the-art person ReID approaches, namely LOMO+XQDA [45], CRAFT [46], GOG+XQDA [47] and WHOS+KISSME [48].

Before showing the performance improvement when the proposed approach is integrated with different state-of-the-art person ReID approaches, we wanted to perform a comparison with other BTF-based approaches first. In order to illustrate the effectiveness of the proposed method in performing color calibration, we first compared the color correlation scores obtained with the proposed approach and cBTF [17] on the VIPeR [1], CUHK01 [35], CUHK02 [52] and CUHK03 [12] datasets. Then, we also performed person ReID evaluation on the VIPeR, CAVIAR4REID, CUHK01 and CUHK03 datasets for comparison with the mBTF, cBTF and wBTF [2].

This experimental results section is organized as follows: First, we describe the different datasets employed in Sec. 2.3.1. The results on color-correlation performance are presented in Sec. 2.3.2. Sec. 2.3.3 summarizes the results of person ReID performance comparison with other BTF-based approaches. Finally, the comparison results, with four different state-of-the-art person ReID approaches, are presented in Sec. 2.3.4.

As proven by the results, the proposed method provides an improved brightness transfer across disjoint camera views and improves ReID performance. For all the experiments and comparisons below, ρ was set to be 0.05, which allowed to have representative BTFs in the codebooks. For the VIPeR dataset, for instance, this resulted in 8, 7 and 7 codewords for R, G and B channels, respectively. 98 codeword triplets were formed (non-zero frequency), and with the proposed ordering approach, they were trimmed down to 32 triplets.

2.3.1 Datasets: VIPeR, CUHK and CAVIAR4REID

In different experiments, we have employed VIPeR, CUHK01, CUHK02, CUHK03 and CAVIAR4REID datasets. The VIPeR dataset contains 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions [1]. Each image is scaled to 128x48 pixels. CUHK01 [35] and CUHK02 [52] datasets have 971 identities with 3884 images and 1816 identities with 7264 images, respectively. CUHK03 [12] has 1360 identities and 13164 images from taken from five disjoint camera pairs. Both manually cropped bounding boxes and automatically detected bounding boxes, obtained using the discriminatively trained part-based models [53], are provided. We have used the autonomously detected bounding boxes in our experiments. CUHK03 dataset includes viewpoint variations, detection errors as well as occlusions. CAVIAR4REID [8] dataset has a total of 72 pedestrians, and 50 of them have both the camera views and the remaining 22 with one camera view, with image size ranging from 17x39 to 72x144.

2.3.2 Color Correlation Comparison

We first divide the dataset into training and test sets. The training set is used to build the codebook of BTFs for the proposed method, and obtain the cBTF for comparison purposes. The images provided in these datasets also include some of the background as seen in Fig. 2.1, 2.4,2.10,2.11 and 2.12, i.e. the persons are not segmented out in the images. We first segmented the persons, as described in Sect. 2.2.1, and obtained the codebook of BTFs and cBTF from the foreground regions.

After training, the transfer functions are applied to the images in the test set coming from one camera, and then the color correlation score is calculated with the images from the other camera. Better brightness transfer will result in higher color correlation scores between images. Since the training set has an effect on the BTFs, we performed 5-fold and 3-fold cross validation for these experiments. More specifically, we divided the dataset as 20% training and 80% testing for 5-fold, and as 33% training and 66% testing for 3-fold, and repeated the separation of dataset randomly five times. For each image pair in the test set, we calculated the color correlation, and averaged it over the total number of image pairs in the test set. Tables 2.1 and 2.2 show the color correlation scores without segmenting out the persons and when persons are segmented out, respectively. As can be seen, segmenting out the persons results in obtaining better codebooks for BTFs, and thus increased color correlation scores. As shown in Table 2.2, for 5-fold validation, the proposed method achieves 20.1%, 16.9%, 18.2% and 14.2% higher color correlation scores than cBTF on VIPeR, CUHK01, CHUHK02 and CUHK03 datasets, respectively. For 3-fold validation, the proposed method achieves 20.3%, 15.2%, 16.7% and 19.4% higher color correlation scores than cBTF on the same datasets. Another observation is that, in general, the proposed method is affected less when the training dataset becomes smaller.

Some example image pairs, and the results of applying the brightness transfer for compensation are presented in Fig. 2.4. Columns (a) and (b) show the images of the same person from first and second camera, respectively. Columns (c) and (d) show the transformed version of the image in column (a) by the proposed method and cBTF, respectively. The red, green and blue channel codewords used by the proposed method are displayed in column (e). As can be seen, the proposed method achieves better compensation and increases the color similarity between different camera views.

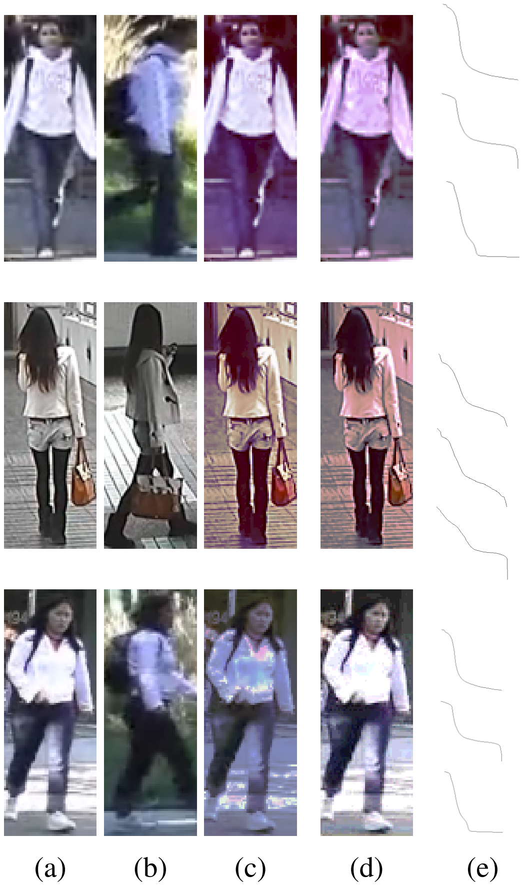


Fig. 2.4: (a) Camera 1 view (b) camera 2 view; (c) and (d) are the compensated or color-transferred version of (a) by the proposed method and by cBTF, respectively; (e) codewords computed by the proposed method for the R,G,B channels (top to bottom).

Dataset	5-folded			3-folded		
	Proposed	cBTF	Impr.	Proposed	cBTF	Impr.
VIPER	**0.2691**	0.2158	24.7%	**0.2957**	0.2481	19.2%
CUHK01	**0.3134**	0.2690	16.5%	**0.3456**	0.3034	13.9%
CHUK02	**0.3279**	0.2812	16.6%	**0.3837**	0.3363	14.1%
CUHK03	**0.3644**	0.3224	13.0%	**0.3691**	0.3147	17.3%

Table 2.1: Improvement in color correlation scores compared to cBTF in VIPER, CUHK01, CUHK02 and CUHK03 datesets before segmenting out persons.

Dataset	5-folded			3-folded		
	Proposed	cBTF	Impr.	Proposed	cBTF	Impr.
VIPER	**0.3494**	0.2909	20.1%	**0.3963**	0.3832	20.3%
CUHK01	**0.4062**	0.3475	16.9%	**0.4554**	0.3954	15.2%
CUHK02	**0.4405**	0.3727	18.2%	**0.4749**	0.4069	16.7%
CUHK03	**0.4081**	0.3521	14.2%	**0.4203**	0.3520	19.4%

Table 2.2: Improvement in color correlation scores compared to cBTF in VIPER, CUHK01, CUHK02 and CUHK03 datesets after segmenting out persons.

2.3.3 Person Re-identification Performance Comparison with BTF-based approaches

As mentioned above, we first compared the person ReID results with the other BTF-based approaches. To show the improvement obtained by the proposed color calibration, and also provide a valid and commensurate comparison to the method in [2] and the other papers included therein, we used the same distance metric between images as in [2], which relies on histograms of RGB, HSV, and YCbCr color features, and histograms of oriented gradients computed for each of the RGB color channels.

We first compared our proposed method with wBTF [2], cBTF and mBTF by using the match rate for different ranks. The results are summarized in the lower half of Table 2.3 for the VIPeR dataset. The table also includes the results for other approaches, namely CPS [8], SDALF [36], PRDC [33], AdaBoost, L1 norm and Bhattacharrya, reported in [2]. For our method, we used the ordering and trimming approach described above, to end up with a smaller number of codeword triplets. The p in the table refers to the number of test images used (out of a total of 632 images in the VIPeR dataset). Thus, the number of training images is $(632 - p)$ for different cases. Rows 10 and 11 show the results obtained with our proposed codebook BTF (CB-BTF) approach, when we employ our proposed background/foreground segmentation method and the one used in [2], respectively. These rows show the improvement obtained when background is eliminated in a better way with the proposed approach. In multi-camera tracking scenarios, segmented foreground regions are obtained after applying background subtraction. Thus, further performance improvement can be achieved. In the table, bold numbers refer to the best results, and our proposed method outperforms all other methods for all scenarios involving different number of test images. Comparing rows 11 and 12, i.e. isolating our proposed FG/BG segmentation approach, shows the effectiveness of the BTFs obtained with the proposed method compared to [2]. In other words, using better BTFs results in improving the target-ReID performance.

We also performed experiments by using all different combinations of codewords exhaustively (rather than trimming the triplets) for Red, Green and Blue channels, and observed that this does not improve the performance significantly. More specifically, when $p = 316$, the obtained rates were 23.82%, 48.34%, 62.09% and 77.78% for $r = 1, r = 5, r = 10$ and $r = 20$, respectively.

Moreover, we used the Cumulative Matching Characteristic (CMC) curves to show the improvement in the person reID performance. CMC curve represents the percentage of images, for which the correct match is in the top r returned images. In CMC curves, the horizontal axis is the r, and the vertical axis is the percentage. In our experiments we randomly divided the dataset into training and test sets 10

	Methods	p=316(# of test set)				p=432				p=532			
		r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20
1	CB-BTF+CRAFT	**52.1**	**81.2**	**89.8**	**95.6**	**43.4**	**75.1**	**86.1**	**90.8**	**35.9**	**66.2**	**80.8**	**87.5**
2	CRAFT[46]	50.3	80.0	89.6	95.5	42.1	73.7	85.2	90.6	35.5	65.8	79.3	86.4
3	CB-BTF+GOG+XQDA	43.4	72.5	82.2	92.1	-	-	-	-	-	-	-	-
4	GOG+XQDA	41.1	71.1	82.1	92.1	-	-	-	-	-	-	-	-
5	CB-BTF + LOMO+XQDA (using RGB,YUV,YCbCr)	42.5	70.2	80.9	91.5	36.4	65.0	76.4	89.9	32.3	61.9	71.2	81.0
6	LOMO+XQDA (using RGB,YUV and YCbCr)	40.9	69.6	80.7	91.5	35.8	64.9	76.4	88.9	31.9	61.7	70.8	80.9
7	LOMO+XQDA [45]	40.0	68.2	80.5	91.1	34.2	63.0	76.4	86.8	29.8	57.9	70.0	80.6
8	CB-BTF+WHOS+KISSME	30.1	60.9	62.1	74.0	24.4	55.9	59.4	68.9	21.0	51.7	56.5	72.0
9	WHOS+KISSME	25.8	58.3	60.1	71.9	22.0	54.7	57.2	68.4	18.8	50.5	55.1	70.1
	Comparison of BTF-based methods												
10	CB-BTF w/prop. seg.	**23.35**	**47.81**	**61.90**	**77.12**	**17.10**	**35.91**	**52.34**	**65.63**	**14.24**	**32.09**	**43.75**	**59.27**
11	CB-BTF w/ seg. in [2]	23.01	47.17	60.58	76.70	16.93	35.84	51.63	64.49	13.75	31.85	43.67	57.69
12	wBTF [2]	21.99	46.84	59.97	75.95	15.05	35.76	50.81	64.24	13.72	31.77	42.86	57.42
13	cBTF	19.27	38.85	53.54	64.69	14.25	31.75	43.96	53.49	12.95	28.76	35.97	46.06
14	mBTF	18.81	38.47	50.90	63.58	14.12	29.70	43.91	52.34	12.63	26.24	33.09	45.75
15	CPS[8]	21.84	46.00	57.21	71.50	-	-	-	-	-	-	-	-
16	SDALF[36]	20.00	38.00	48.50	65.00	-	-	-	-	-	-	-	-
17	PRDC[33]	15.66	38.42	53.86	70.09	12.64	31.97	44.28	59.95	9.12	24.19	34.40	48.55
18	AdaBoost	8.16	24.15	36.58	52.12	6.83	19.81	29.75	43.06	4.19	12.95	20.21	37.73
19	L1-Norm	4.18	11.65	16.52	22.37	3.80	9.81	13.94	19.44	3.55	8.29	12.27	17.59
20	Bhattacharyya	4.65	11.49	16.55	23.83	4.19	10.35	14.19	20.19	3.82	9.08	12.42	17.88

Table 2.3: Results on the VIPeR dataset showing the percentage of images for which the correct match is in the top r returned images. p is the number of images in the test set. Bold numbers represent the best-score in every column.

times, and used the average of 10 trials to obtain the CMC curves. We performed a comparison with wBTF [2] on both CAVIAR4REID and VIPeR datasets, and plotted the CMC curves. Figures 2.5(a) and 2.5(b) show the CMC curves obtained on VIPeR and CAVIAR4REID, respectively. For VIPeR, the training dataset size was 100, and for CAVIAR4REID the number of training images was only 25. As can be seen, the proposed method outperforms wBTF, and the improvement in performance is more pronounced when the number of training images gets smaller.

We also plotted the CMC curves for the proposed method and wBTF, with and without segmenting out the persons from the background. The results are shown in Fig. 2.6. As can be seen, the proposed approach is less sensitive to not segmenting out the persons. Fig. 2.7 shows example matching results for $r = 1$, obtained with the proposed method and wBTF.

In addition to the VIPeR dataset, we used CUHK01 and CUHK03 datasets for person ReID performance comparison of the proposed method with other BTF-

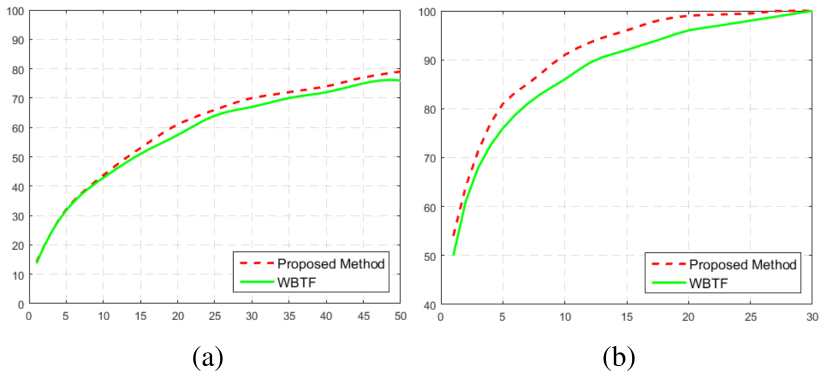


Fig. 2.5: CMC curves for the proposed method and WBTF [2] for (a) VIPeR (p = 532, # of training images is 100) and (b) CAVIAR4REID (# of training images is 25) datasets.

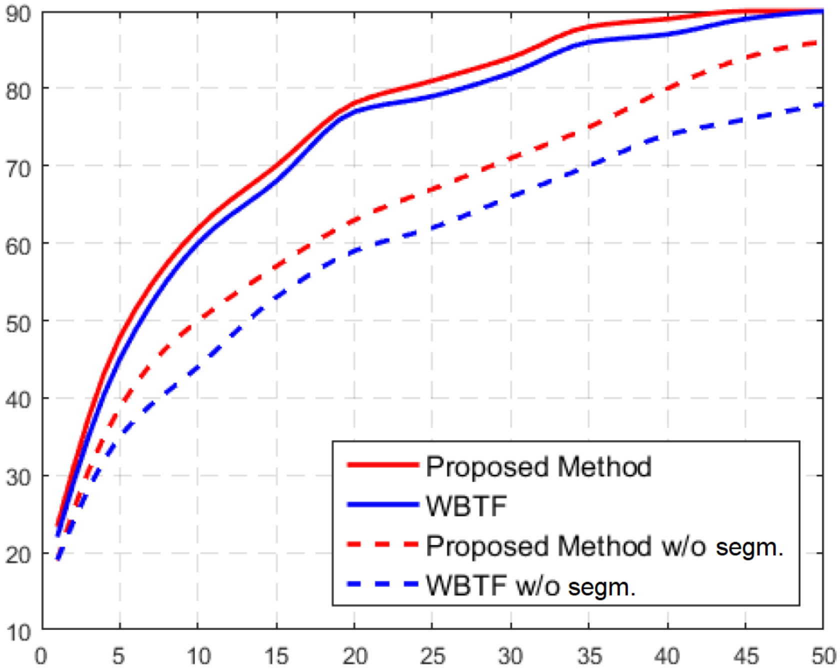


Fig. 2.6: Performance comparison of the proposed method and wBTF, with and without segmenting out the persons from the background, on VIPeR dataset with p = 316.

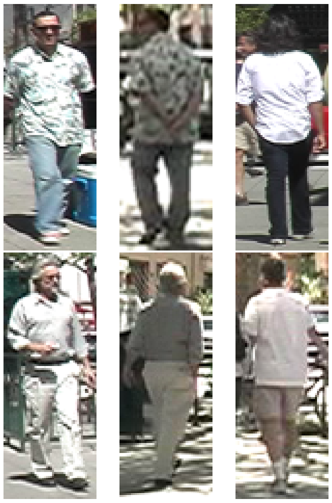


Fig. 2.7: Sample rank 1 ReID results. (a) Query image, (b) rank 1 matching by proposed method, (c) rank 1 matching by wBTF.

based approaches, as well as more recent state-of-the-art approaches. The results are summarized in Tables 2.4 and 2.5, respectively, and will be discussed more in Sect. 2.3.4.

2.3.4 Person Re-identification Performance Comparison with Recent State-of-the-art Methods

In order to support our argument that different target ReID approaches can benefit from our proposed approach and further extend the generalization of our proposed method, we have performed extensive experiments with four different state-of-the-art person ReID approaches, namely LOMO+XQDA [45], CRAFT[46], KISSME [48] and GOG+XQDA [47]. We used VIPER, CUHK01 and CUHK03 datasets for performance comparison.

In [45], the extracted features are color description (YUV histogram) and Scale Invariant Local Ternary Pattern (SILTP). A subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) is also presented in [45]. In our comparison experiments, we first performed brightness compensation

Table 2.4: CUHK01 results

Dataset	CUHK 01			
Rank	1	5	10	20
CB-BTF+CRAFT	**80.1**	**93.0**	**95.3**	**97.8**
CRAFT	74.5	91.2	94.8	97.1
CB-BTF+LOMO+XQDA (using RGB,YUV,YCbCr)	65.4	85.2	91.1	95.0
LOMO+XQDA(RGB,YUV,YCbCr) (using RGB,YUV,YCbCr)	64.1	83.9	90.2	94.7
LOMO+XQDA	63.2	83.6	90.2	94.7
CB-BTF+GOG+XQDA	60.3	79.1	86.4	92.1
GOG+XQDA	57.2	77.9	86.2	92.1
CB-BTF+WHOS+KISSME	31.3	59.0	73.2	86.8
WHOS+KISSME	29.4	57.7	72.4	86.1
CB-BTF w/prop. seg.	15.01	28.98	42.40	65.85
CB-BTF w/ seg. in [2]	13.42	26.37	41.29	65.17
WBTF	10.93	22.19	35.51	57.83
CBTF	9.62	20.34	33.86	56.39
MBTF	8.73	19.75	33.72	55.52

Table 2.5: CUHK03 results

Dataset	CUHK 01			
Rank	1	5	10	20
CB-BTF+CRAFT	**85.6**	**97.2**	**98.3**	**99.1**
CRAFT	84.3	97.1	**98.3**	**99.1**
CB-BTF+LOMO+XQDA	54.7	83.8	92.3	96.3
LOMO+XQDA(RGB,YUV,YCbCr)	53.4	82.9	92.2	96.3
LOMO+XQDA	52.2	82.2	92.1	96.3
CB-BTF+GOG+XQDA	69.2	92.3	96.4	97.2
GOG+XQDA	67.3	91.0	96.0	97.2
CB-BTF+WHOS+KISSME	16.6	50.0	53.9	71.8
WHOS+KISSME	14.2	48.5	52.6	70.5
CB-BTF w/prop. seg.	12.31	23.60	38.75	59.86
CB-BTF w/ seg. in [2]	10.48	21.56	37.33	59.09
WBTF	9.29	16.84	27.24	57.06
CBTF	8.73	14.52	26.76	53.79
MBTF	8.52	12.41	24.56	52.82

by using our constructed codebook for R, G and B channels as shown in Fig. 2.8. We then performed feature extraction for LOMO+XQDA by using only the YUV channels. In this case, the improvement was not significant. Then, we added RGB and YCbCr channels into the color description, and expectedly, the performance improved for all rank results. For instance, as seen in Tables 2.3, 2.4 and 2.5, rank 1 results improved by 1.6%, 2.2% and 2.5% for VIPeR, CUHK01 and CUHK03 datasets, respectively.

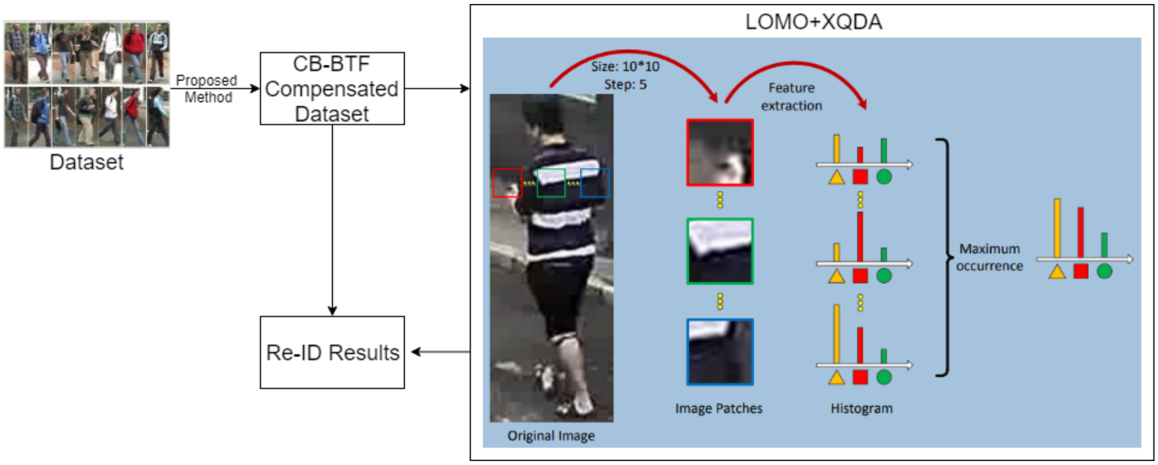


Fig. 2.8: Experiments with LOMO framework.

Another state-of-the-art method we studied is CRAFT [46], which is a deep learning-based approach. We used the source code that is available from the project web site [54]. Similar to above steps, we first performed brightness compensation by using our constructed codebook for R, G and B channels, and then applied CRAFT on the compensated images as shown in Fig. 2.9. As seen in Tables 2.3, 2.4 and 2.5, rank 1 results improved by 1.8%, 5.6% and 1.3% for VIPER, CUHK01 and CUHK03 datasets, respectively.

The third method we used in our experiments is GOG+XQDA [47]. Similar to the above results, applying brightness compensation with our proposed method before using the person ReID approach improved results for all ranks and all datasets. As seen in Tables 2.3, 2.4 and 2.5, rank 1 results improved by 2.3%, 3.1%

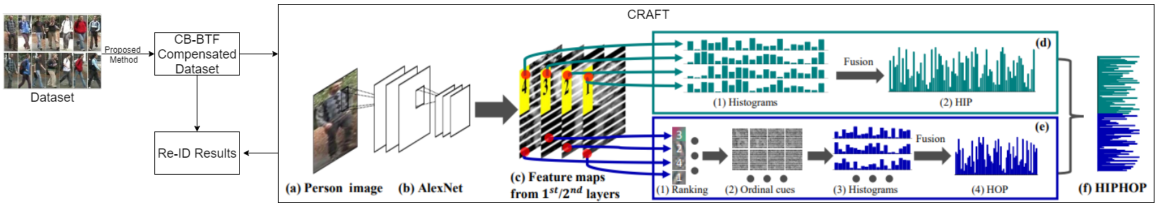


Fig. 2.9: Experiments with CRAFT framework.

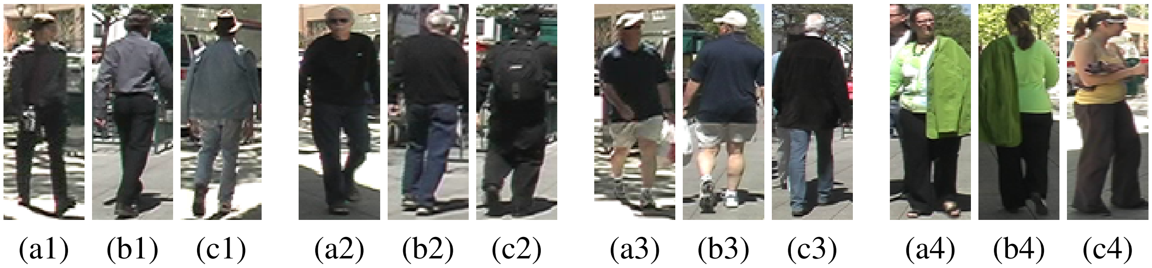


Fig. 2.10: Example Rank 1 ReID results on the VIPeR dataset. (a1)-(a4) query images, (b1) CB+CRAFT (c1) CRAFT; (b2) CB+GOG (c2) GOG; (b3) CB+LOMO (c3) LOMO; (b4) CB+KISSME (c4) KISSME.

and 1.9% for VIPER, CUHK01 and CUHK03 datasets, respectively.

The fourth method we used in our experiments is WHOS+KISSME [48]. When our method is used for brightness compensation beforehand the rank 1 performance improves by 4.3% on the VIPeR dataset. Since the performance of WHOS+KISSME was much lower on this dataset compared to CRAFT, GOG+XQDA and LOMO+XQDA, we have not performed a test for WHOS+KISSME on CUHK01 and CUHK03 datasets.

Figures 2.10, 2.11 and 2.12 show different example Rank 1 results, on VIPeR, CUHK01 and CUHK03 datasets, respectively. These examples illustrate the cases, where a base person ReID approach returns the wrong person as rank 1, whereas using our proposed method beforehand for brightness compensation results in the correct match in rank 1.

2.4 Conclusion

We have presented a novel approach to better model the appearance variation across disjoint camera views, and improve the performance of any person re-

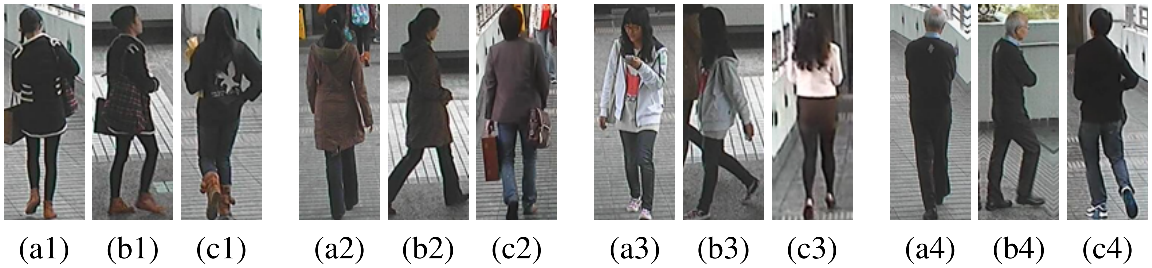


Fig. 2.11: Example Rank 1 ReID results on the CUHK01 dataset. (a1)-(a4) query images, (b1) CB+CRAFT (c1) CRAFT; (b2) CB+GOG (c2) GOG; (b3) CB+LOMO (c3) LOMO; (b4) CB+KISSME (c4) KISSME.

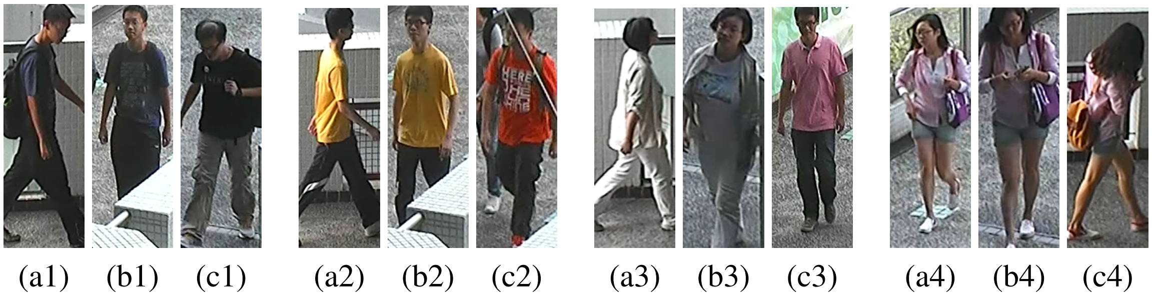


Fig. 2.12: Example Rank 1 ReID results on the CUHK03 dataset. (a1)-(a4) query images, (b1) CB+CRAFT (c1) CRAFT; (b2) CB+GOG (c2) GOG; (b3) CB+LOMO (c3) LOMO; (b4) CB+KISSME (c4) KISSME.

identification approach that incorporates color/brightness histograms and appearance models. We have proposed building a codebook of brightness transfer functions, and also an ordering and trimming criteria to increase computational efficiency. We have performed extensive set of experiments on different commonly used datasets. Results have shown that the proposed method outperforms other BTF-based approaches. Moreover, the proposed approach was incorporated into four different state-of-the-art person re-identification methods, and an increased top rank matching rate has been obtained for all methods and on all datasets, supporting our initial argument above.

Chapter 3

Person Detection and Re-identification Across Multiple Images and Videos Obtained via Crowdsourcing

3.1 Introduction

Large camera networks are increasingly being deployed in public places such as airports, subways, campuses and office buildings. These cameras are usually installed across wider areas with non-overlapping fields of view (FOV) to provider better coverage. However, with these setups, tracking of person(s) is still limited to the areas where the cameras are installed. Thanks to the widespread use of smart phones, crowdsourcing of videos has the potential to provide a much larger coverage for longer periods of time. For instance, a video or image captured by someone at a remote location (with no pre-installed camera) can have the person of interest in it. If it is possible to have access to large number of videos captured at different locations over time, this would potentially allow following a person over a much larger area for extended periods of time. However, the extremely large amount of data captured in these scenarios makes it impossible for humans to perform manual analysis, and necessitates autonomous analysis of data. Video analysis enable long term characterization and modeling of the people in the scene. Such application is required for high-level surveillance tasks and make them even smarter [19].

Given an image/video of person of interest, re-identification is the process of identifying same person in the images/video captured by a different camera. Re-identification is indispensable in establishing the consistent labeling across multiple cameras or even within the same camera to connect broken trajectories or re-establish lost tracks. Person re-identification is a difficult problem due to large variations in person's appearance, lighting conditions and contrast across different cameras. Person re-identification/association across different camera views has two main parts; (1) detecting the people in the scene and (2) re-identifying them in other camera views. The face detection method proposed by Viola and Jones [55], and the human detection method, based on histograms of oriented gradients, proposed by Dalal and Triggs [56] are two of the important works on detection. Dollar et al. [57] proposed Integral Channel Features for pedestrian detection. Local Binary Patterns [58], multiple kernels [59] and part-based models [53] have also been introduced for human detection.

Deep Convolutional Neural Networks (CNNs) have received a lot of attention recently, especially after achieving very good performance in the ImageNet challenge [60]. Later, Girshick et al. [61] combined region proposals with CNNs, and introduced R-CNN, regions with CNN features, for object detection. Then, faster R-CNN [62] has been proposed, which focus on the speed up by pruning the hypothesis in detection.

Person re-identification has been studied by many researchers in the past few years [63, 64, 65, 66, 67, 68, 69, 70]. Earlier research focused on taking advantage of camera inner parameters during the matching process. Most of the existing work relies on the appearance-based similarity between images, such as color and texture of clothing, to establish correspondences. In general, recent approaches focus on three aspects; (1) designing subject-discriminative [71], descriptive and robust visual descriptors to characterize a person's appearance [72], (2) using feature transformation which projects features between different camera-dependent spaces, such as feature warping [73], and sparse basis expansion [74, 75], and (3) learning suitable distance metrics that maximize the chance of a correct matching [76, 77, 4].

In many of the existing works, cameras are static and background subtraction is applied to detect moving objects. This significantly simplifies and speeds up the detection stage. Then, focus is placed on the matching part. However, if cameras are mobile or only single images (not videos) are available, then background subtraction cannot be employed. In this case, human detection needs to be performed in the entire image, which is in general a computationally intensive process. In this work, different from most of the existing work, we focus on a crowdsourcing scenario to find and follow person(s) of interest in the collected images/videos. We propose a novel approach combining R-CNN based person detection with the GPU implementation of color histogram and SURF-based re-identification. Moreover, the GeoTags are extracted from the EXIF data of videos captured by smart phones, and these locations are displayed on a map together with the time-stamps for a spatio-temporal representation/visualization of the trajectory. All the processing, including R-CNN based detection, histogram correlation and SURF-based matching, is performed on GPU. The average processing time for the proposed detection and matching algorithm is 5 ms per frame on an NVIDIA Quadro K5200

8GB GPU. The experimental results show the promise of the proposed method.

The rest of the chapter is organized as follows: The proposed approach is described in detail in Sec. 3.2. The experimental results are presented in Sec. 5.4, and the chapter is concluded in Sec. 5.5.

3.2 Proposed Method

In our proposed method, we use the pre-trained Region-based Convolutional Networks (R-CNN) [61], implemented with Caffe [78], to detect people in images. We refer to the image containing the person(s) of interest as the query image and the frames in the video(s) to be examined as the candidates images. The R-CNN detector, which is run on all video images, returns the bounding boxes of candidate people regions in each image. Then, both the color histogram and SURF features are extracted from the candidate regions to be matched with the subject(s) of interest. After a match is found, GeoTag information is extracted from the matched video to generate the spatio-temporal model for candidates, and mark their path on a map. The overall flow diagram of the proposed method is provided in Fig. 3.1.

3.2.1 R-CNN-based Detection

We apply the R-CNN model [61] provided in Caffe for better performance in detection results. A higher detection rate will provide better matching results across different views. In our experiments, we observed that all people in the videos were successfully detected by the R-CNN detector.

The architecture of the used R-CNN is shown in Fig. 3.2. It has 7 layers. Since the training images for the convolutional neural network needs to be a fixed-size of 227×227 pixels, all the images are resized to the required size first. As stated in [79], in the first layer, the resized images are convolved with 96 kernels of size

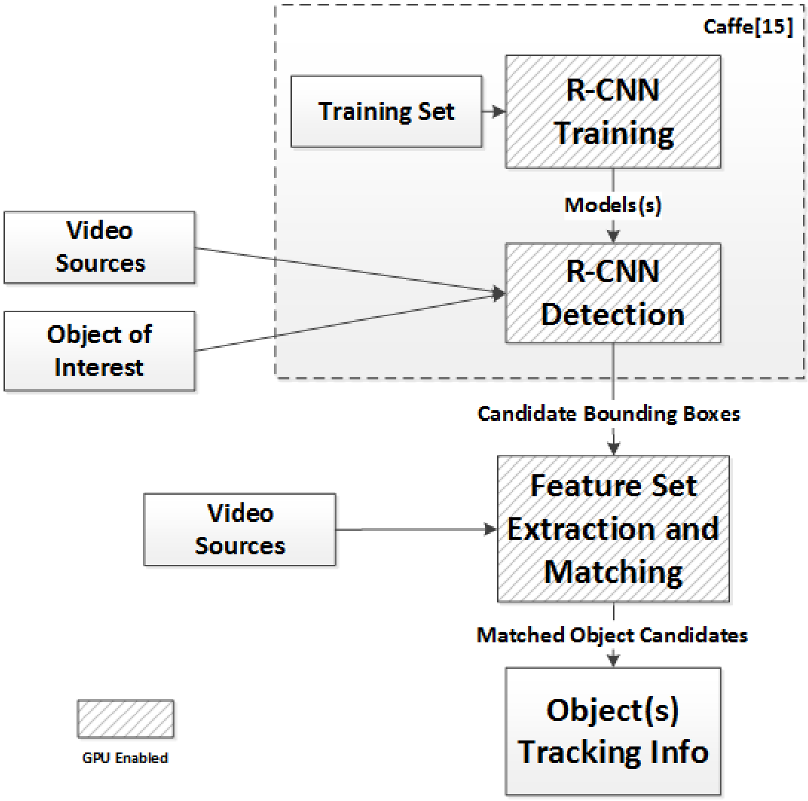


Fig. 3.1: Flow diagram of the proposed method.

11x11x3 pixels with a stride of 4 pixel, and then max-pooling is applied in 3x3 grid. The second layer has the same framework, with 256 kernels of size 5x5x48. Layer 3 and 4 are two convolution layers without pooling, which both have 384 kernels. Layer 5 is similar to layer 2. Layer 6 and 7 are fully connected layers with 4096 nodes. The activation function used in convolution and fully connected layer is Rectified Linear function. More details can be found in [60] and [61].

3.2.2 GPU Accelerated Re-identification

The process of detection and re-identification of people across many different views can be computationally very expensive. Each region proposed by the R-CNN on each candidate image has to be compared with the representation of person(s) of interest in the query image to see if there is a match. In order to reduce the pro-

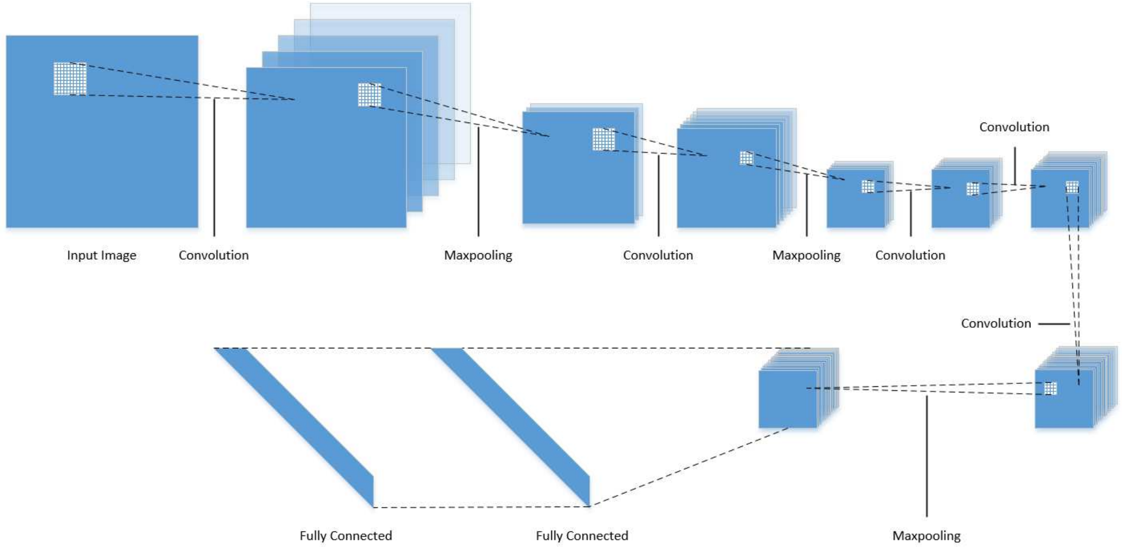


Fig. 3.2: Architecture of used R-CNN model.

cessing time of each video frame, we implemented a GPU-based parallel matching strategy, which speeds up the detection and re-identification phase. The matching procedure has the following steps:

1. Apply R-CNN based detection on video frames to get bounding boxes for candidate people regions.

2. Convert candidate regions to HSV color space, compute histograms and compare with the histogram of the person(s) of interest from the query image by using color correlation.

3. If the histogram correlation score is higher than a threshold τ, apply SURF to detect matching feature points between candidate regions and images person(s) of interest.

As mentioned previously, R-CNN detection provides the coordinates of the bounding box around candidate people regions. Then, we convert this region to the HSV color space to reduce the effect of illumination changes. Afterwards, we calculate the color histograms, and compare the histograms by using color correlation calculation in (3.1), where H_1 and H_2 are the two color histograms, N is the total number of histogram bins, and $\overline{H} = \frac{1}{N}\sum_i H(i)$.

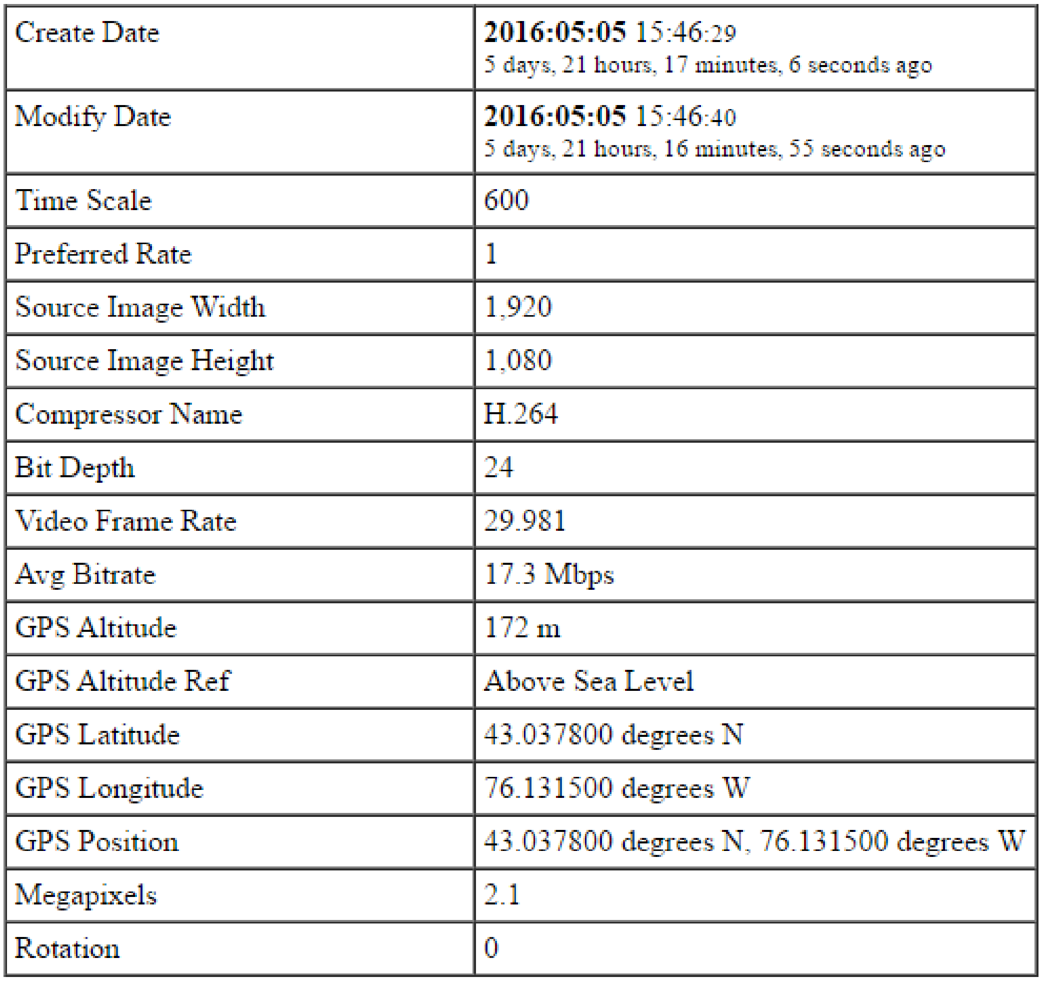
$$d(H_1, H_2) = \frac{\sum_{i=1}^{N}\left(H_1(i) - \overline{H_1}\right)\left(H_2(i) - \overline{H_2}\right)}{\sqrt{\sum_{i=1}^{N}\left(H_1(i) - \overline{H_1}\right)^2\left(H_2(i) - \overline{H_2}\right)^2}}. \tag{3.1}$$

If the histogram correlation score is higher than a threshold τ, a matcher based on Speeded Up Robust Features (SURF) [80] is applied to detect matching feature points between candidate regions and images person(s) of interest. SURF is a fast, scale- and rotation-invariant detector and descriptor, which outperforms previous methods with respect to repeatability, distinctiveness and robustness. In this work, a SURF-based detector and matcher is implemented on GPU to recognize the same person from multiple different views. If the number of matched points is larger than 50% of the smaller number of points (between the candidate region and the person of interest), a match is declared.

3.2.3 Spatio-Temporal Model

The prevalence of smart phones not only provides large volumes of video/image data with ever-increasing quality, but also makes GPS information, via GeoTags, more accessible. To generate the spatio-temporal model of a person, who is matched across different videos captured by different phones at different locations, we extract the EXIF information, and then visualize it on GOOGLEMAP API. EXIF is an exchangeable file format that contains the metadata embedded within the video. As shown in Fig 3.3, it includes information such as the length of the video, GPS location, timestamp and compression rate. We use EXIFtool [81] to extract the GeoTags, and obtain the location where the video was initially captured.

It should be noted that the GeoTags only provide the GPS location of the device



Create Date	**2016:05:05** 15:46:29 5 days, 21 hours, 17 minutes, 6 seconds ago
Modify Date	**2016:05:05** 15:46:40 5 days, 21 hours, 16 minutes, 55 seconds ago
Time Scale	600
Preferred Rate	1
Source Image Width	1,920
Source Image Height	1,080
Compressor Name	H.264
Bit Depth	24
Video Frame Rate	29.981
Avg Bitrate	17.3 Mbps
GPS Altitude	172 m
GPS Altitude Ref	Above Sea Level
GPS Latitude	43.037800 degrees N
GPS Longitude	76.131500 degrees W
GPS Position	43.037800 degrees N, 76.131500 degrees W
Megapixels	2.1
Rotation	0

Fig. 3.3: Exif information from video

when it starts recording the video. Thus, we need to estimate the distance between the target and the device to display the location of the target. We do this by using the bounding box sizes for people obtained at four different pre-defined distances from smart phones. Table 3.1 provides the list of bounding box sizes obtained at 20 ft, 40 ft , 60 ft and 80 ft from the capturing smart phone. We find the closest bounding box size from this list and use the corresponding distance as the distance of the target from the phone. Then, we draw a circle with this radius around the GPS location of the device. Since the videos are obtained through crowdsourcing, there can be another video of the same person captured around the same time. We

Distance from camera (ft)	Bounding box size
20	300 x 950
40	143 x 454
60	100 x 283
80	68 x 220

Table 3.1: Bounding box sizes for different distances from the camera.

follow the same steps for the location from another device, and use the intersection of the circles as the location of the target. This process is illustrated in Fig. 3.4.

As mentioned above, the GPS location is only available for the first frame of the video. Thus, to be able to track the location of a target over time, we need many videos captured at different times, and this is when crowdsourcing comes to the rescue. When there are enough videos obtained through crowdsourcing, it will provide a more complete spatio-temporal picture of the trajectory of the target. It should be noted that, the proposed approach can be applied to images as well as videos. Since, each image has its own GPS and time-stamp information, localization will be more precise compared to videos. To be able to use the intersection of the circles as the location, when we have videos as the input, we used short videos in our experiments (since GPS data is only available for the first frame).

3.3 Experimental Results

As mentioned above, all the processing, including R-CNN based detection, histogram correlation and SURF-based matching, is performed on GPU. The average processing time for the proposed detection and matching algorithm is 5 ms per frame on an NVIDIA Quadro K5200 8GB GPU.

We performed different experiments using videos captured by multiple smart phones at different places. The targets are seen from different views (frontal, side view etc.) in the captured videos. In the first experiment, a total of three videos

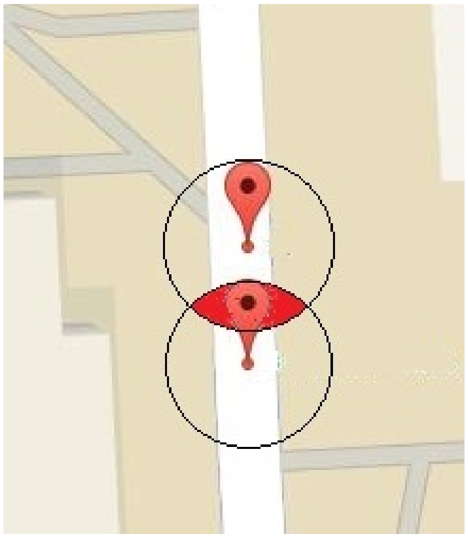
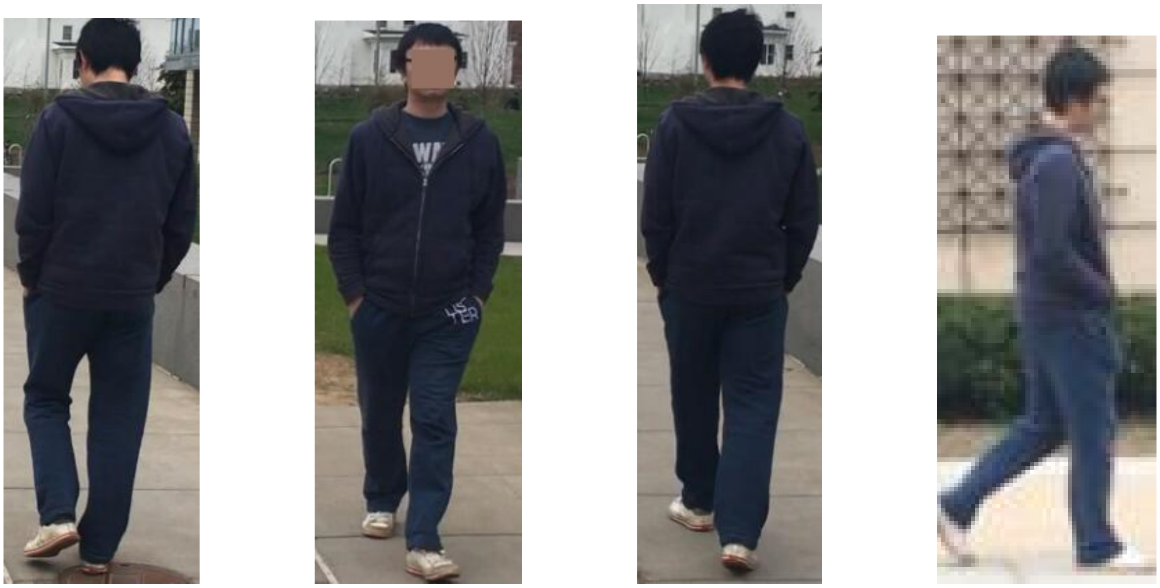


Fig. 3.4: Determining the location of a target.

were recorded, from three different phones, capturing a target from different angles, namely front view, side view and back view. Figure 3.5 shows the query target image and the matched images, from different cameras, showing the target from different angles. The color histogram correlation scores are shown in Table 3.2. As expected, since the query is a back view, the correlation between the query and a candidate region with the back view is the highest. The correlation score for the front view comes next, and the score between the query and a side view is the smallest for the same person. This experiment also allowed us to determine an empirical threshold for τ that allows matching the same person from different views, and also having a discriminative power for different people. The GeoTag locations on a map, for this experiment, are displayed in Fig. 3.6.

Experiment 1	**Correlation score**
Front view	0.6875
Back view	0.7124
Side view	0.4992

Table 3.2: Color histogram correlation scores obtained between back, and front and side views in the first experiment.



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

Fig. 3.5: Detection and matching results to the query image in (a). Matched candidates from (b) front view, (c) back view, and (d) side view.

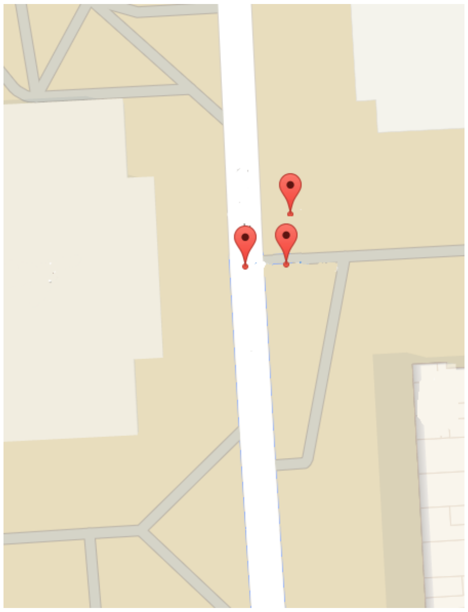


Fig. 3.6: The GeoTag locations displayed on a map for the first experiment.

The second experiment scenario has a more complicated setup involving the detection and matching of three different targets simultaneously. There are four different smart phones capturing videos of these three targets from different locations and angles. To better simulate a crowdsourcing scenario, a total of 38 videos are recorded from these four devices so that there are more videos to process, which would provide a more complete GeoTag information over time. Each video contains at least one target or more.

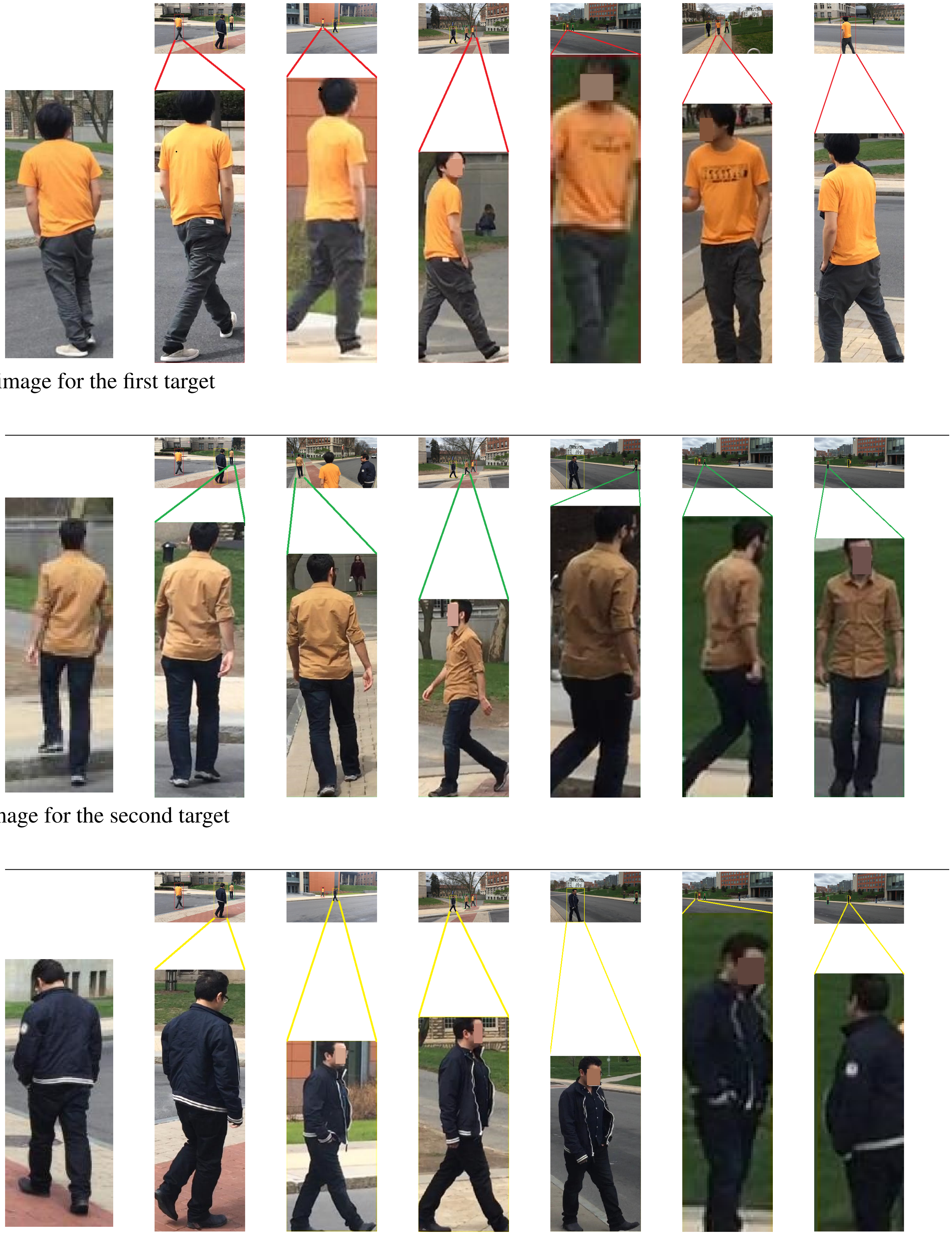
Example matching images are shown in Fig. 3.7. There is a large variation in the distance of targets from the capturing device, which can be observed from the original frames as well as the lower resolution of some of the detected targets. All three candidates have been successfully re-identified across different cameras, capturing from different angles, by histogram correlation and SURF-based matching. Table 3.3 shows example color correlation scores between the query image and all three candidates. As can be seen, the highest correlation scores are across the diagonal. Fig. 3.8 shows the GeoTag locations on a map for all three targets over time. Red, green and yellow marks correspond to targets one, two and three, respectively.

Color correlation	candidate 1	candidate 2	candidate 3
Query 1	0.6847	0.3975	0.1875
Query 2	0.4434	0.6685	0.2349
Query 3	0.1524	0.2569	0.7176

Table 3.3: Color histogram correlation scores between the query and candidate regions for the second experiment.

3.4 Conclusion

We have proposed a method, which performs R-CNN based person detection and then person re-identification via matching on GPU. A pre-trained model from R-CNN is used to detect person candidate regions in videos. Then, both color and SURF features are extracted for each candidate, and used for matching with the person of interest. The GPS location in the image/video EXIF information is used to obtain a spatio-temporal model for the path taken by the target, and these locations are displayed on a map. All the processing is performed on GPU, and takes an average of 5ms to process one frame. As feature work, we will incorporate more features to improve the re-identification.



(a) Query image for the first target

(b) Query image for the second target

(c) Query image for the third target

Fig. 3.7: Matching results for the experiment involving the detection and matching of three different targets simultaneously.

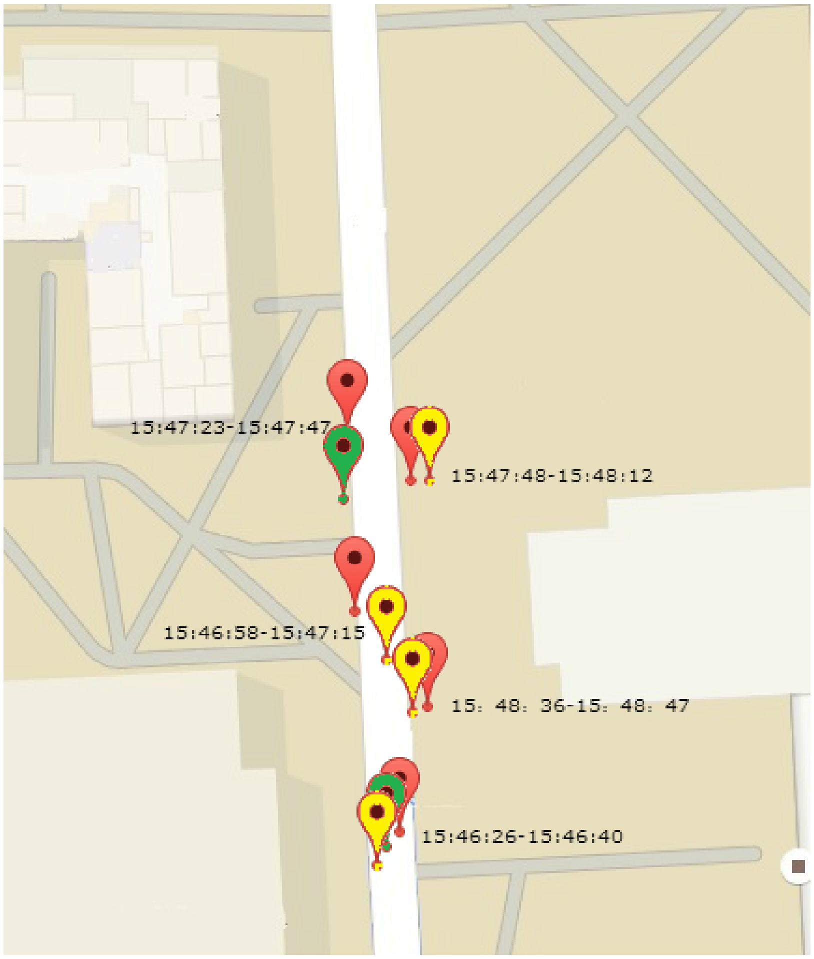


Fig. 3.8: GeoTag results on the map for the second experiment.

CHAPTER 4
ADVERSARIAL ATTACKS ON PERSON RE-IDENTIFICATION IN VIDEO SURVEILLANCE

4.1 Introduction

In order to continuously track targets across multiple cameras with disjoint views, it is essential to re-identify the same target across different cameras. However, this is a very challenging task due to several reasons including changes in illumination and target appearance, and variations in camera intrinsic parameters and viewpoint.

There has been great interest and significant progress in person re-identification (ReID) [82, 83, 84, 85, 86, 87], which is important for security and wide-area surveillance applications as well as human computer interaction systems. Fueled by the new models, including the neural network-based approaches, proposed in recent years, the performance of person ReID approaches has improved significantly. For instance, the rank-1 accuracy of the state-of-the-art method on the Market 1501 dataset [88] is 94.8% [82], which has increased from 44.4% when the dataset was initially released in 2015.

In this chapter, we demonstrate the effectiveness of an attack model in generating adversarial examples (AEs) for the person ReID application, attack multiple state-of-the-art person ReID models, and also compare the performance of the presented attack approach with other state-of-the-art attack models via an extensive set of experiments on various person ReID benchmark datasets. One of our goals is to demonstrate the extreme vulnerability of multiple state-of-the-art person ReID approaches to this attack, and draw the attention of the research community to the existing security risks. In person ReID, the paired probe and gallery images are expected to have high similarity. However, by adding human-imperceptible perturbations to the probe images, the models are easily fooled even when the probe images appear the same as the original images.

Adversarial examples [20, 21, 22] have been extensively investigated recently in image classification [22, 23], object detection [24, 25, 26] and semantic segmentation [27, 24], etc. However, relatively less attention has been paid to the robustness of person ReID models. Bai et al. [89] proposed an adversarial metric attack, which targets on fooling the distance metrics in person ReID systems. An early attempt for the defense shows that a metric-preserving network can be applied to defend against such attack. Zheng et al. [90] propose Opposite Direction Feature Attack (OFDA) to generate adversarial examples/queries for retrieval tasks such as person ReID. The idea is to push away the feature of the adversarial query in the opposite direction of the original feature.

In this chapter, we present and employ an effective approach to generate adversarial examples targeting person ReID methods. Our approach [91] is referred to as the Dispersion Reduction (DR), and it is a black-box attack. The main idea behind our approach is reducing the “contrast" of the internal feature map of a neural network. The intuition is that just like reducing the contrast of an image would make the objects less recognizable or distinguishable, reducing the contrast of an internal feature map would have a similar effect on recognizability of objects by the neural network. In our previous work [91], we showed the transferability of the DR attack across different tasks including object detection, classification and text recognition. The contribution of this work includes the following: We adapt the DR attack for the person ReID problem, and perform an extensive comparison and evaluation on different state-of-the-art methods and multiple benchmarks. In addition, we compare the performances of multiple attack methods. We show that making a feature map “featureless", through dispersion reduction, is very-well suited to fool any state-of-the art ReID model. Moreover, we use different network models (different from the model used by the victim ReID networks) as the source model to generate the adversarial examples and show the effectiveness and generalizability of our attack approach. We also analyze the effect of the perturbation budget on the attack performance.

The rest of the chapter is organized as follows: The related works on both person ReID and attack models are summarized in Section 4.2. The proposed dispersion reduction-based attack approach, and the methodology are described in Section 4.3. The experimental results are presented in Section 5.4, and the chapter is concluded in Section 4.5.

4.2 Background and Related Work

4.2.1 Review of Person ReID Methods

Various person re-identification (ReID) approaches have been proposed in the past [3], which can be classified into different categories. There have been methods based on distance learning [33, 34, 35, 4, 5, 6, 7], on feature design and selection [36, 8, 37, 38, 9, 39, 10], and on mid-level feature learning [40, 41, 11, 12, 13].

Many works relied on color transformation and statistic models for person re-identification. Cheng and Piccardi [42] applied a cumulative color histogram transformation and employed an incremental major color spectrum histogram representation. Trajectory matching, height estimation and illumination-tolerant color representation were used by Madden and Piccardi [43]. Chae and Jo [44] employed a Gaussian Mixture Model (GMM) for the segmented regions in a person, and used a ratio of the GMMs to identify the same person. The Brightness Transfer Function (BTF) and its variants have been introduced to improve the matching performance. Porikli [15] proposed BTF for inter-camera color calibration. Later, Javed et al. [16] and Posser et al. [17] proposed the Mean Brightness Transfer Function (mBTF) and the Cumulative Brightness Transfer Function (cBTF), respectively. Datta et al. [2] presented the Weighted-BTF (wBTF), and Bhuiyan et al. [18] presented the Minimum Multiple Brightness Transfer Function (Min-MCBTF) to model the appearance variation by using a learning approach. However, it was assumed that multiple consecutive images are available for training, which is not the case for the commonly used benchmark datasets.

Researchers then focused on combining the features and distance metrics at the same time. Liao et al. [45] proposed Local Maximal Occurrence (LOMO) and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) for person ReID. Chen et al. [46] formulated a new view-specific person ReID framework, referred to as camera correlation aware feature augmentation (CRAFT). In this framework, cross-view feature adaptation is performed by measuring cross-view correlation from visual data distribution and carrying out adaptive feature augmentation. Matsukawa et al. [47] proposed the hierarchical Gaussian Of Gaussian (GOG) descriptor, which generates discriminative and robust features that describe color and textural information simultaneously. An image is first divided into horizontal strips. Then, local patches in the strips are modeled using a Gaussian distribution. Köstinger et al. [48] proposed the KISSME, which is a statistical inference perspective to address the problem of metric learning.

More recent works employ neural networks and achieve state-of-the-art performance in person ReID. Zheng et al. [82] proposed DG-Net that encompasses a generative module, which separately encodes a specific person into both appearance and structure. It also integrates a discriminative module that shares the appearance encoder with the generative module. As a result, the high-quality cross-id composed images are fed back to the appearance encoder online and used to improve the model for discriminative module. Zhang et al. [84] proposed AlignedReID performing automatic part alignment during learning, without requiring extra supervision or pose estimation. By learning jointly on global and local features, it aims to address existing drawbacks. Xie et al. [92] proposed PLR-OSNet, which introduces Part-level resolution (PLR) into Omni-Scale Network (OSNet) [93]. It has two branches including both global and local feature representations. The global branch adopts a global-max-pooling layer, while the local branch employs a part-level feature resolution scheme for producing only a single ID-prediction loss, which is in contrast to existing part-based methods.

4.2.2 Adversarial Attack Methods

Szegedy et al. [21] introduced the adversarial images, which can fool the Convolutional Neural Network (CNN)-based models, and cause misclassification by adding small perturbations to the original images. In one of the earlier works, Goodfellow et al. [94] proposed fast gradient sign method (FGSM), which generates AEs in one step. Several works extended this by iteratively updating the AEs with multi-step attacks including the basic iterative method (BIM) [22], deep fool [95], momentum iterative method [23], Diverse Inputs Method (DIM) [96] and Translation-Invariant (TI) attacks [97]. Compared with FGSM, the iterative methods generate a smaller perturbation, which makes the adversarial examples even more imperceptible to human eye.

The transferability property of adversarial examples motivated research on black-box adversarial attacks. To perform black-box attacks, methods have been introduced [98, 99], which employ a substitute model that is trained to mimic the target model. Gradient-free attacks use feedback on query data, $i.e.$, soft predictions [100, 101] or hard labels [102]. However, these aforementioned approaches require feedback from the target model, which is not practical in some scenarios. More recently, several methods have been proposed, which study the attack generation process itself. In general, an iterative attack [20, 103, 29] achieves a higher attack success rate than a single-step attack [94] in a white-box setting, but performs worse when transferred to other models. Below, we will summarize some of these attack methods.

Gradient-based Adversarial Attack Methods

Fast Gradient Sign Method (FGSM) [94] generates the adversarial example x^{adv} by linearizing the loss function in the input space and performing one-step update as follows

$$x^{adv} = x^{real} + \epsilon \cdot sign(\nabla_x J(x^{real}, y)), \tag{4.1}$$

where $\nabla_x J(x^{real}, y)$ is the gradient of the loss function w.r.t. x, $sign(\cdot)$ is the sign function that constrains the perturbation in L_∞ norm bound. FGSM can generate more transferable adversarial examples, however, it may not be as effective in white-box attacks [22].

Basic Iterative Method (BIM) [22] extends FGSM by updating the gradient in a multi-step manner with a small step size α, which can be expressed as

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\nabla_x J(x_t^{adv}, y)), \tag{4.2}$$

where $x_0^{adv} = x^{real}$. BIM clips x_t^{adv} after each update, or sets $\alpha = \epsilon/T$, with T being the number of iterations to ensure that they are in an ϵ-neighbourhood of the real image.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [23] integrates momentum term into the iterative attack process. The update procedure is as follows

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y)}{\left\|\nabla_x J(x_t^{adv}, y)\right\|_1}, \\ x_{t+1}^{adv} &= x_t^{adv} + \alpha \cdot sign(\nabla_x J(x_t^{adv}, y)), \end{aligned} \tag{4.3}$$

where g_t collects the gradient information up to the t-th iteration, and μ is the decay factor.

Diverse Inputs Method (DIM) [96] applies random and differentiable transformations to the input images with probability p and maximizes the loss function with respect to these transformed inputs. The transformed images are fed into the classifier for gradient calculation. Such transformation includes random resizing and padding with a given probability p. This method can be combined with the momentum-based method to further improve the transferability.

Translation-Invariant Attack Methods

Translation-Invariant (TI) attack methods have been proposed by Dong et al. [97] to further improve the transferability on white-box models. The authors notice the difference between the discriminative regions used by defenses to identify object categories and the normal trained models. Rather than optimizing the objective function, TI attack method uses a set of translated images to optimize the adversarial examples as

$$\begin{gathered}\arg\max_{x^{adv}} \sum_{i,j} w_{ij} J(T_{ij}(x^{adv}, y)), \\ s.t. \left\| x^{adv} - x^{real} \right\|_\infty \leq \epsilon\end{gathered} \tag{4.4}$$

where $T_{ij}(x)$ is the translation operation that shifts image x by i and j pixels along the two-dimensions, respectively, and w_{ij} is the weight for the loss $J(T_{ij}(x^{adv}, y))$.

Note that the TI can be integrated into any gradient-based attack such as FGSM or DIM. For example, the translation-invariant method for fast gradient sign method (TI-FGSM) updates as

$$x^{adv} = x^{real} + \epsilon \cdot sign(\mathbf{W} * \nabla_x J(x^{real}, y)), \tag{4.5}$$

Also, the translation-invariant method for diverse inputs method (DIM) can also be obtained by a similar approach.

4.3 Proposed Approach

In this section, we will describe the dispersion reduction-based attack on the person ReID application.

4.3.1 Notation

We use x^{real} to denote the original query image, and $f(\cdot)$ to denote a deep neural network classifier. The output feature map at layer k is denoted by $\mathfrak{F}$, where $\mathfrak{F} = f(x^{real})|_k$ at the first time step. For each step afterwards, we calculate the dispersion, which is denoted as $g(\cdot)$, and the gradient of dispersion as $\nabla_{x^{real}} g(\mathfrak{F}_k)$ to update the adversarial examples x^{adv}. More details will be provided in the following section.

4.3.2 Dispersion Reduction

For person ReID, the existing models are trained with various benchmark datasets, which have different labeling schemes. Thus, compared to the image classification problem, person ReID is more complicated. More specifically, treating and attacking the person ReID models as black-boxes require an approach that is highly transferable and effective at attacking different training datasets and model architectures. The aforementioned existing black-box attacks, however, use a pre-trained model as surrogate, which shares the same training dataset and same labeling scheme with the targeted models. Moreover, most existing attack methods rely on task-specific loss functions, which greatly limits their transferability across tasks and different network models.

In our previous work [91], we showed that Dispersion Reduction (DR) has good transferability properties, and is successful in across task attack scenarios. DR employs a publicly available classification network as the surrogate source model, and attacks models that are used in different computer vision tasks, such as object detection, semantic segmentation and cloud API applications. DR is a black-box attack. Conventional black-box attacks establish a source model as the surrogate, for which the inputs are paired with the labels generated from the target model instead of the ground truth labels. In this way, the source model mimics the behavior of the target model. Our proposed DR attack, on the other hand, does not rely on the labeling system or a task-specific loss function, since DR only accesses top part of the model. Although a source model is still required, there is no need for training with new target models or querying the target model for labels. Instead, a pre-trained public model could simply serve as the source model due to the strong transferability of the proposed DR attack. As shown in Fig. 4.1, the DR attack reduces the contrast of an internal feature map, by reducing its dispersion, so that the information that is in the feature map becomes indistinguishable, and

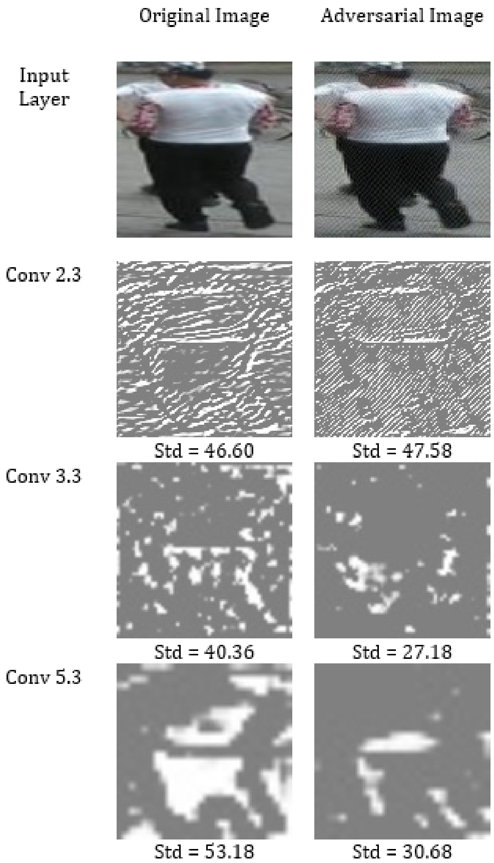
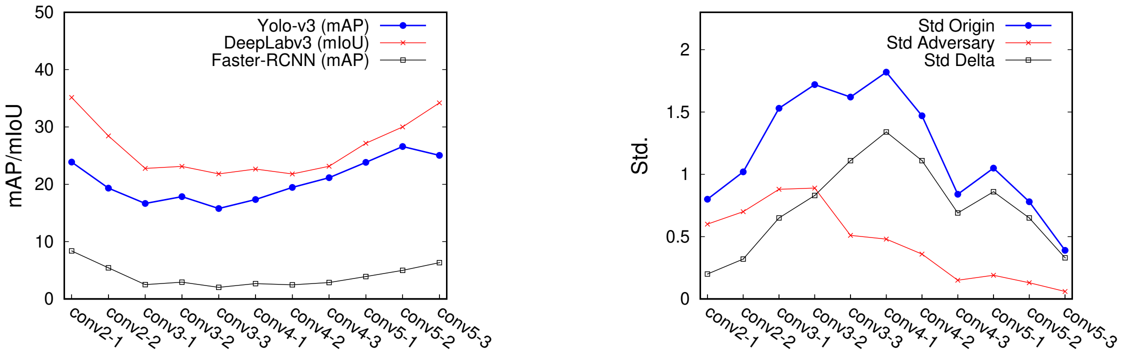


Fig. 4.1: The DR attack reduces the dispersion of an internal feature map. This adversarial example was generated by attacking (reducing the dispersion of) the conv3.3 layer of VGG16 model, which also results in the distortion of the feature maps of the subsequent layers (e.g. conv5.3) compared to the original image feature maps.

the following layers are not able to extract any useful information regardless of what kind of computer vision task is at hand. The adversarial example, shown in the second column of Fig. 4.1, was generated by attacking (reducing the dispersion of) the conv3.3 layer of VGG16 surrogate model. This also results in the distortion of the feature maps of the subsequent layers (e.g. conv5.3). As can be seen, compared to the feature maps of the original image, the standard deviations of the feature maps for the adversarial image are lower after the attacked layer.

Moreover, we have analyzed the effect of attacking different convolutional layers of the VGG16 network with the proposed DR attack based on the PASCAL VOC2012 validation set [91]. Fig. 4.2a shows the mAP value for Yolov3 and Faster



(a) mAP/mIoU results.

(b) Std. before and after attack

Fig. 4.2: Effect of the DR attack when different layers of VGG16 is attacked. Attacking the middle layers results in higher drop in the performance compared to attacking the top or bottom layers.The drop in standard deviation of middle layers is also larger than the top and bottom layers.

RCNN, and mIoU for Deeplabv3 and FCN. Fig. 4.2b is the plot of the standard deviation values before and after the DR attack, together with the change. As can be seen, attacking the middle layers of VGG16 results in higher drop in the performance compared to attacking top or bottom layers. At the same time, the change in the standard deviation for middle layers is larger compared to the top and bottom layers. We can infer that for initial layers, the budget ϵ constrains the loss function to reduce the standard deviation, while for the layers near the output, the standard deviation is already relatively small, and cannot be reduced too much further. Based on this observation, we choose one of the middle layers as the target of the DR attack. More specifically, in our experiments, we attack conv3-3 for VGG16, the last layer of group A for inception-v3, and the last layer of 2nd group of bottlenecks (conv3-8-3) for ResNet152.

The DR attack is defined as the following optimization problem:

$$\begin{aligned} &\min_{x} g(f(x^{adv}, \theta)) \\ s.t.\ &\left\|x^{adv} - x^{real}\right\|_{\infty} \leq \epsilon, \end{aligned} \tag{4.6}$$

where $f(\cdot)$ is a deep neural network classifier, θ denotes the network parameters, and $g(\cdot)$ computes the dispersion. As shown in Alg. 2, our proposed DR takes iterative steps that create adversarial examples by reducing the dispersion of an intermediate feature map at layer k. Dispersion describes the extent to which a distribution is stretched or squeezed, and there can be different measures of dispersion such as variance, standard deviation, and gini coefficient [104]. We have chosen to use the standard deviation (denoted by $g(\cdot)$) as the dispersion metric due to its simplicity.

Given any feature map, DR iteratively adds perturbation to x^{real} along the direction of decreasing standard deviation, and maps it to the vicinity of x^{real} by clipping at $x \pm \epsilon$. Denote the feature map at layer k as $\mathfrak{F} = f(x_t^{adv})|_k$, DR attack solves the following equation

$$\begin{aligned} x_{t+1}^{adv} &= x_t^{adv} - \nabla_{x^{adv}} g(\mathfrak{F}_k) \\ &= x_t^{adv} - \frac{dg(t)}{dt} \cdot \frac{df(x_t^{adv})|_k}{dx^{adv}}. \end{aligned} \tag{4.7}$$

The code is provided in [105].

Algorithm 2 Dispersion Reduction Attack

```
    Input : classifier $f$, real image $x^{real}$, feature map at layer $k$, perturbation $\epsilon$, iteration $T$
and learning rate $l$
    Output: adversarial example $x^{adv}$,
        $s.t. \left\| x^{adv} - x^{real} \right\|_\infty \leq \epsilon$
procedure DISPERSION REDUCTION
    $x_0^{adv} \leftarrow x^{real}$
    for t=0 to T-1 do
        $\mathfrak{F}_k = f(x_t^{adv})|_k$
        Compute std $g(\mathfrak{F}_k)$
        Compute gradient $\nabla_{x^{real}} g(\mathfrak{F}_k)$
        Update $x^{adv}$ by:
        $x_t^{adv} = x_t^{adv} - Adam(\nabla_{x^{real}} g(\mathfrak{F}_k), l)$
        Project $x_t^{adv}$ to the neighbour of $x^{real}$:
        $x_{t+1}^{adv} = clip(x_t^{adv}, x^{real} - \epsilon, x^{real} + \epsilon)$
    end forreturn $x_{t+1}^{adv}$
end procedure
```

4.3.3 Victim ReID Models and Implementation Details of Attacks

In order to evaluate the effectiveness of our proposed adversarial DR attack, we adapt it for the person ReID problem, and attack three different state-of-the art person ReID appraoaches, namely DG-Net [82], AlignedReID [84] and PLR-OSNet [92]. For person re-identification, both DG-Net and AlignedReID use ResNet-50 [106] as the backbone model, while PLR-OSNet employs the Omni-Scale Network as the backbone. DG-Net reaches 94.8% and 86.0% on the rank-1 accuracy and mean average precision (mAP), respectively, on Market 1501 dataset [88]. AlignedReID achieves 92.6% and 82.3% accuracy [107] on the rank-1 and mAP, respectively,, and PLR-OSNet achieves 95.6% and 88.9% accuracy on the rank-1 and mAP, respectively, on the Market 1501 dataset.

We used the pre-trained models for these ReID approaches, provided by the authors on their Github pages [108, 109, 110]. During training, the images are resized to 256$\times$128, which is a strong baseline that can achieve higher accuracy. We reduce the mini-batch size from 16 to 4 to save GPU memory usage on all models and all datasets. The learning rate for DG-Net, AlignedReID and PL-OSNet is 0.0001, 0.0002 and 0.0003, respectively. All models use a decay rate of γ = 0.1, which reduces the learning rate by a factor of 1/10 after T steps during the training. For DG-Net T is set to 60000. For AlignedReID and PL-OSNet, T is set to and 20, respectively. More implementation details can be found in the source codes provided by the authors [108, 109, 110].

For each dataset, the images are separated into training and testing folders. We follow the data preparation process described in [108, 109]. After pre-processing, we apply the TI-FGSM and TI-DIM attacks as described in [97], and detailed in the source code on Github page [111].

For our dispersion reduction (DR) attack, we first used the pre-trained ResNet-152 as the source model. The values of the parameters, listed in Algorithm 2, are as follows: $\epsilon = 4$, l (learning rate)= 0.05, $T = 100$. The adversarial examples are generated on the test images, and used for testing on the victim ReID models. As mentioned above, both DG-Net and AlignedReID use ResNet-50 [106] as the backbone model. Thus, in order to generate the adversarial examples with different surrogate models, we have also used VGG-16 and InceptionV3, as our source models. As discussed above, we used conv3-3 for VGG16, the last layer of group A for inception-v3, and the last layer of 2nd group of bottlenecks (conv3-8-3) for ResNet152, as the attack layers. We also analyzed the effects of using different ϵ values, and a detailed discussion is provided in the following section.

4.4 Experiments, Results and Discussion

As mentioned above, we have used three state-of-the-art ReID methods as victim models, attacked them with the proposed DR attack, and evaluated the performance drop on four different datasets. Moreover, we attacked the same victim models with two other state-of-the-art attack approaches, namely TI-FGSM and TI-DIM [97, 111]. We compared the effectiveness of our DR attack with these other attack methods as well. Moreover, we have used three different network models as the surrogate source model to evaluate and compare the performance drop and the attack effectiveness.

4.4.1 Datasets

We have employed four challenging and commonly used benchmark datasets to demonstrate the effectiveness of the proposed attack. These datasets are Market-1501 [88], CUHK 03 [12], DukeMTMCreID [112] and MSMT 17 [113], which are briefly described below.

Market-1501

Market-1501 [88] dataset contains 32,217 images of 1501 labeled persons from six camera views. There are 751 identities in the training set and 750 identities in the testing set. In the original study on this proposed dataset, mAP is the evaluation criteria used to compare the algorithm performances.

CUHK03

CUHK03 [12] dataset contains 8765 images of 1467 labeled persons. In this chapter, we use a new protocol, in which the training set and test set have 767 and 700 identities, respectively. We select the detected bounding boxes instead of labeled bounding box results. It is a more difficult evaluation protocol for CUHK 03.

DukeMTMC-ReID

DukeMTMC-ReID [112] dataset is composed of 36,411 images of 1812 persons captured from eight cameras. There are 702 identities in the training set and 1110 identities in the testing set. The evaluation criteria is mAP, same with the Market-1501 dataset.

MSMT17

MSMT17 [113] is the largest image-based person ReID dataset introduced in 2018. It contains 124,069 labeled images of 4101 person IDs captured from 12 different outdoor or indoor cameras. The evaluation protocol/criteria is also same as the Market-1501 dataset, and uses mAP.

Approch	Market1501	CUHK 03	DukeMTMC-reID	MSMT 17
DG-Net (DG) [82]	86.0	61.1	74.8	52.3
AlignedReid (AR) [84]	82.3	70.7	82.8	43.7
PLR-OSNet(OS)[92]	88.9	77.2	81.2	–
TI-FGSM-DG [97, 94, 82]	51.2	31.9	48.3	32.1
TI-FGSM-AR [97, 94, 84]	58.7	29.4	49.5	30.8
TI-FGSM-OS [97, 94, 92]	55.9	32.1	51.2	-
TI-DIM-DG [97, 96, 82]	45.4	14.2	32.7	25.4
TI-DIM-AR [97]	44.9	16.5	29.4	22.6
TI-DIM-OS [97, 96, 92]	47.7	19.1	35.6	-
DR-DG (Proposed)	**28.9**	**7.8**	**20.2**	**15.3**
DR-AR (Proposed)	**20.2**	**8.3**	**12.3**	**12.8**
DR-OS (Proposed)	**29.3**	**9.5**	**19.6**	**-**

Table 4.1: mAP values of victim models, before and after attacks, on different datasets (ϵ = 4). Row 1 to 3 are the results from state-of-the-art baseline methods. Row 4-12 are the performances after attacked by different attack approaches. Lower numbers mean better attack performance.

4.4.2 Evaluation Metric

With the same image perturbation ($\epsilon = 4$), we compare performances of all the attack methods while attacking the victim ReID approaches. The lower number indicates more drop in ReID accuracy, and thus, better attack performance. Mean average precision (mAP) is used as the evaluation metric. The effects of using different ϵ values are discussed in Section 4.4.4.

4.4.3 Results and Discussion

In the first set of experiments, we used ResNet-152 as the source model of the attacks. The results are summarized in Table 4.1, wherein the first three rows show the mAP values for the baseline victim models, namely DG-Net, AligneDReID and PLR-OSNet, on different benchmark datasets. The mAP values for DG-Net are 86.0%, 61.1%, 74.8% and 52.3%, and the mAP values for AlignedReID are 82.3%, 70.7%, 82.8% and 43.7% for Market1501, CUHK03, DukeMTMC-ReID and MSMST17 datasets, respectively. For PLR-OSNeT, the mAP values are 88.9%, 77.2% and 81.2% for Market1501, CUHK03 and DukeMTMC-ReID datasets, respectively. These three models are regarded as the state-of-the-art ReID approaches based on their performance. The fourth to sixth rows in Table 4.1 show the value of the mAP after the victim models are attacked with TI-FGSM, which is a state-of-the-art attack method. For instance, for the Market-1501 dataset, the mAP value of DG-Net drops by 34.8 from 86.0 to 51.2, the mAP value of AlignedReID drops by 23.6 from 82.3 to 58.7, and the mAP value of PLR-OSNet drops by 33 from 88.9 to 55.9. For the CUHK03 dataset, the mAP value of DG-Net drops by 29.2 from 61.1 to 31.9, the mAP value of AlignedReID drops by 41.3 from 70.7 to 29.4, and the mAP value of PLR-OSNet drops by 45.1 from 77.2 to 32.1. The seventh to ninth rows in Table 4.1 show the value of mAP after the victim models are attacked with TI-DIM, which is another state-of-the-art attack method. Compared to TI-FGSM, this attack is more effective since it causes more drops in the mAP values for all four datasets. For instance, for the CUHK03 dataset, the mAP value of DG-Net drops by 46.9 from 61.1 to 14.2, the mAP value of AlignedReID drops by 54.2 from 70.7 to 16.5, and the mAP value of PLR-OSNet drops by 58.1 from 77.2 to 19.1. The last three rows of Table 4.1 show the mAP value after the victim models are attacked with the proposed DR approach. As can be seen, our proposed approach is the most effective attack compared to TI-FGSM and TI-DIM, and causes the most drop in the mAP values for all victim models and for all four datasets. For instance, for the CUHK03 dataset, the mAP value of DG-Net drops by 53.3 from 61.1 to 7.8, the mAP value of AlignedReID drops by 62.4 from 70.7 to only 8.3, and the mAP value of PLR-OSNet drops by 67.7 from 77.2 to only 9.5.

Fig. 4.3 shows some example images and query results for the Market-1501 dataset. The first column shows the query images, and columns 2 through 11 show the Rank 1 to Rank 10 returned images for that query, respectively. The first and third rows are for the original query images, while the second and fourth rows are for the adversarial query images. The perturbations between the query images of first versus second row and third versus fourth row are imperceptible to the human eye, but the person ReID performances have been significantly impacted by the proposed attack. Similar results for CUHK03 and DukeMTMC-ReID datasets are shown in Fig. 4.4 and Fig. 4.5, respectively. We report the overall results for the MSMT17 dataset in Table 4.1, and are not able to provide example images due to the release agreement.

The examples in Figures 4.3, 4.4 and 4.5 show the effectiveness of the proposed DR attack. In these figures, the adversarial examples, although imperceptible to human eye, result in no matches even in Rank 10 returns. As a quantitative measure, we computed the peak signal to noise ratio (PSNR) as well as the structural similarity index measure (SSIM) between the adversarial images (generated by the TI-FGSM, TI-DIM and the proposed DR attack) and the original images, and calculated the average on the Market1501 dataset. The average SSIM value is 0.70, 0.72 and 0.72 for TI-FGSM, TI-DIM and the DR attacks, respectively. The average PSNR is 26, 28 and 27 for TI-FGSM, TI-DIM and the DR attacks, respectively. Since the perturbation budget is kept the same ($\epsilon = 4$) for all the attack methods, their average SSIM and PSNR values are similar. Some example adversarial images generated by these attacks are shown in Fig. 4.6 for qualitative comparison.

In the second set of experiments, we used two other network models, namely VGG-16 and InceptionV3, as our surrogate source models. The goal here was to use different network models, other than ResNet, to generate adversarial examples and show the generalizability of the proposed DR approach. We have generated AEs by using these different networks as the source models with the proposed DR approach and with TI-DIM. We then used the AEs to attack DG-Net and AlignedReID. In this experiment, we chose to use TI-DIM, since it has better attack perfor-

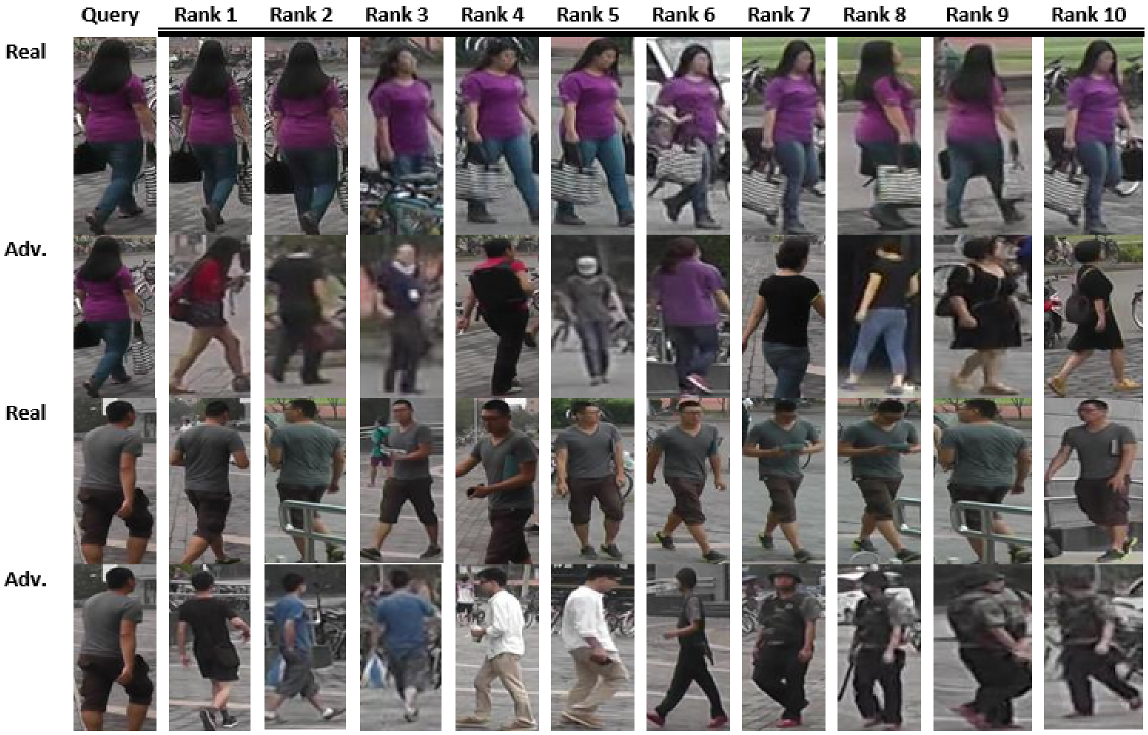


Fig. 4.3: Example query results on the Market-1501 dataset. First column: Query images; columns 2 through 11: Rank 1 through Rank 10 returned images, respectively. Rows 1 and 3 are for the original images, and rows 2 and 4 are for the adversarial query images.

mance than TI-FGSM based on Table 4.1. The results obtained with our proposed DR attack are summarized in Tables 4.2 and 4.3 when the victim ReID method is AlignedReID and DG-Net, respectively. As can be seen, when Resnet-152 is used as the surrogate model, it results in the highest drop in the mAP values. This is mostly because most of the ReID approaches use ResNet as their backbone network. However, even when we use VGG-16 or InceptionV3 as the surrogate source model, the proposed DR attack still causes significantly more drop in the mAP values compared to the state-of-the-art attack methods when they use ResNet as their surrogate model (please see Tables 4.1, 4.2 and 4.3).

The results obtained with the TI-DIM are summarized in Tables 4.4 and 4.5, which show the results obtained with the TI-DIM attack, with using different surrogate models, when the victim ReID method is AlignedReID and DG-Net, respectively. When we compare Table 4.2 with Table 4.4, and Table 4.3 and Table 4.5, it can

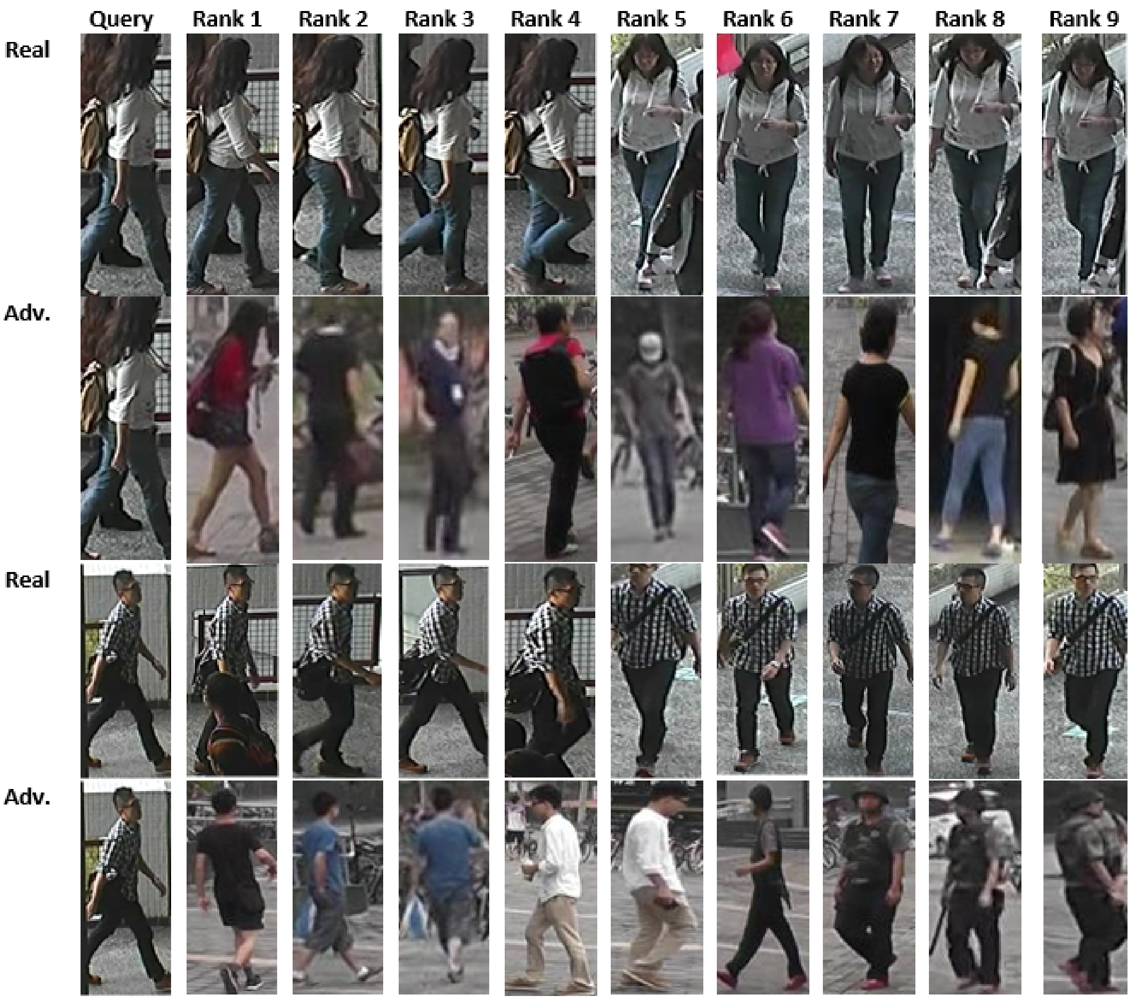


Fig. 4.4: Example query results on the CUHK03 dataset. First column: Query images; columns 2 through 10: Rank 1 through Rank 9 returned images, respectively. Rows 1 and 3 are for the original images, and rows 2 and 4 are for the adversarial query images.

be seen that the proposed DR attack still outperforms the TI-DIM as a black-box attack even when the surrogate model is different from the target model.

Victim	Market 1501	CUHK 03	DukeMTMC-reID	MSMT 17
AlignedReID	82.3	70.7	82.8	43.7
Source netwrk				
InceptionV3	22.5	9.9	14.6	16.2
VGG-16	23.1	9.6	14.5	14.3
Resnet-152	20.2	8.3	12.3	12.8

Table 4.2: mAP values on different datasets when AlignedReID is attacked with the proposed DR approach. First row is the performance before attack. Last three rows show the results when AEs are generated by using different network models as the surrogate models.

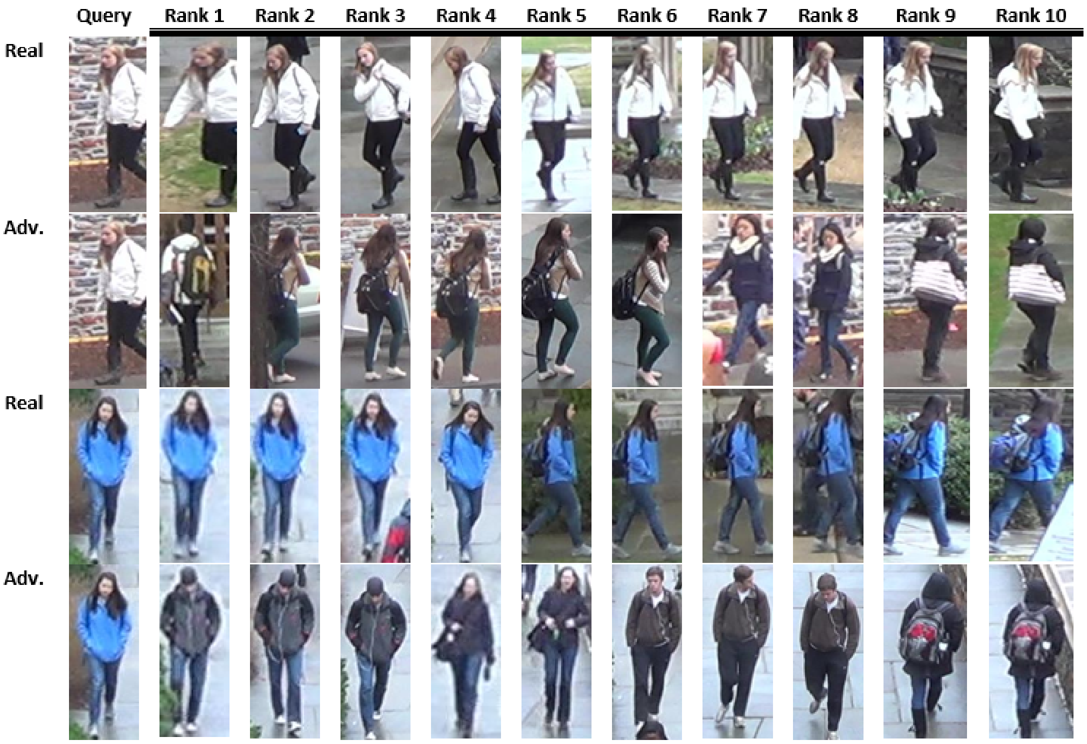


Fig. 4.5: Example query results on the DukeMTMC-ReID dataset. First column: Query images; columns 2 through 11: Rank 1 through Rank 10 returned images, respectively. Rows 1 and 3 are for the original images, and rows 2 and 4 are for the adversarial query images.

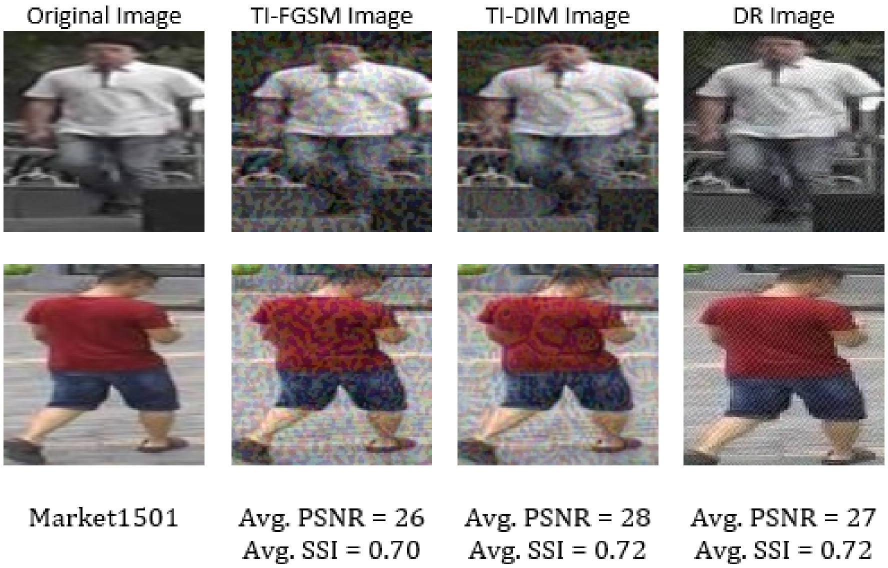


Fig. 4.6: The average PSNR and SSI values (between the original and adversarial images) on the Market1501 dataset for different attack methods when $\epsilon = 4$.

Victim	Market 1501	CUHK 03	DukeMTMC-reID	MSMT 17
DG-Net	86	61.1	74.8	52.3
Source netwrk				
InceptionV3	35.2	8.9	23.2	17.6
VGG-16	31.7	8.6	22.8	18.5
Resnet-152	28.9	7.8	20.2	15.3

Table 4.3: mAP values on different datasets when DG-Net is attacked with the proposed DR approach. First row is the performance before attack. Last three rows show the results when AEs are generated by using different network models as the surrogate models.

TI-DIM	Market 1501	CUHK 03	DukeMTMC-reID	MSMT 17
AlignedReID	82.3	70.7	82.8	43.7
Source netwrk				
InceptionV3	48.2	19.1	35.7	28.1
VGG-16	49.4	18.6	36.2	27.9
Resnet-152	44.9	16.5	29.4	22.6

Table 4.4: mAP values on different datasets when AlignedReID is attacked with the TI-DIM. First row is the performance before attack. Last three rows show the results when AEs are generated by using different network models as the surrogate models.

TI-DIM	Market 1501	CUHK 03	DukeMTMC-reID	MSMT 17
DG-Net	86.0	61.1	74.8	52.3
Source netwrk				
InceptionV3	55.6	20.1	42.4	29.5
VGG-16	57.1	19.5	40.8	29.7
Resnet-152	45.4	14.2	32.7	25.4

Table 4.5: mAP values on different datasets when DG-Net is attacked with the TI-DIM. First row is the performance before attack. Last three rows show the results when AEs are generated by using different network models as the surrogate models.

4.4.4 Effect of ϵ on the performance

In literature, it is a common practice to fix the value of ϵ, and then compare the performance degradation for different attack methods. In the experiments above, we set $\epsilon = 4$, since it results in less change in the original image, and better demonstrates the difference between the attack methods. When ϵ is increased, more budget is given to each attack method to make changes on the original images, and they start to provide similar performance. A better attack should be able to provide more performance degradation with a smaller ϵ budget. As shown in Table 4.6

and Fig. 4.7, our proposed DR attack can reach a given attack effectiveness by using the least budget. For instance, the proposed DR attack drops the mAP value of DG-Net to 20.3 with an ϵ budget of 8, whereas TI-DIM needs a budget of 12 to drop the mAP to 21.8.

Methods/mAP	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 16$
TI-DIM-DG	62.5	45.4	32.5	21.8	12.9
TI-DIM-AR	64.2	44.9	29.4	21.0	18.5
DR-DG	47.4	28.9	20.3	13.5	12.3
DR-AR	48.1	20.2	21.8	15.9	11.2

Table 4.6: mAP values obtained with different ϵ values on Market-1501 dataset while attacking DG-Net and AlignedReID with TI-DIM and the proposed DR attack.

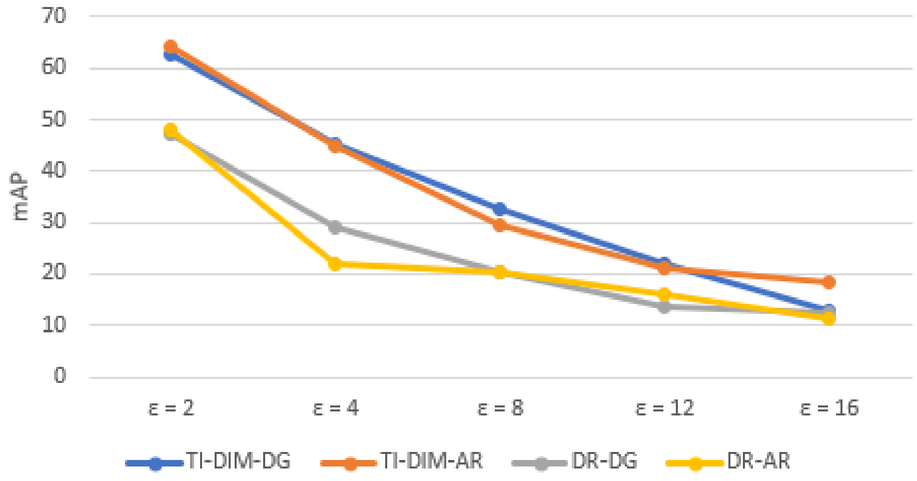


Fig. 4.7: Effect of using different perturbation budget (ϵ) on the attack performance.

4.5 Conclusion

Neural network-based methods have achieved state-of-the-art performance on the person re-identification problem across different camera views. In this chapter, we have presented a black-box and effective attack model, which is based on dispersion reduction, and does not rely on task-specific loss functions and label queries. We have used the adversarial examples generated by this approach to attack three different state-of-the-art person Re-ID models. We have also compared the performance of our attack approach with two other state-of-the-art attack models. The results demonstrate the effectiveness and generalizability of the proposed dispersion reduction attack on three state-of-the-art person ReID models. It also outperforms other state-of-the-art attack models by a large margin, and results in the most drop in the mean average precision values.

Chapter 5
Part-based Feature Squeezing to Detect Adversarial Examples in Person Re-identification Networks

5.1 Introduction

Person re-identification (ReID) describes the task of finding a person from a gallery of images given a probe image of the same person. It is commonly used for tracking a person across different camera views. There has been great interest and significant progress in person ReID, which is very important for security and wide-area surveillance. With the introduction of the neural network-based approaches in recent years, the performance of person ReID has improved significantly. For instance, the mean Average Precision (mAP) of the state-of-the-art method on Market 1501 dataset is 95.5%, which has increased from 44.4% when the dataset was initially released in 2015.

In the past few years, researchers have shown the vulnerability of DNNs against adversarial examples. Despite the impressive performance achieved thanks to DNNs, neural network-based ReID methods also inherit their vulnerability. Adversarial examples, which are carefully crafted images with perturbations that are imperceptible to human eyes, can drastically degrade the performance of most DNN-based ReID approaches. In the meantime, defense approaches against adversarial examples have been proposed [28, 29, 30, 31, 32]. However, defending ReID networks against adversarial attacks is still relatively unexplored.

In this chapter, we present a new method to detect adversarial examples presented to a person ReID network by utilizing part-based feature squeezing. Feature squeezing was proposed for detecting the adversarial examples in image classification task with efficient computation compared to other iterative methods [32]. We show that by applying the feature squeezer on top of the body parts-based ReID model, the AE detection performance can be further improved compared to using a network that is not based on parts. We also show that by detecting AEs, the mAP of person ReID models can be increased compared to not detecting AEs at all. With the PCB model, the mAP after AE detection can reach closeto 70% (compared to an mAP of 22.6 before AE detection).

5.2 Related Work

5.2.1 Person ReID

Different person ReID approaches have been proposed over the years. There have been methods based on hand-crafted features, using attributes like appearance or pose, and part-based features. In recent years, the best performances have been achieved by DNN-based approaches. Zheng et al. [82] proposed the DG-Net containing a generative module, which separately encodes a specific person into both appearance and structure. It also integrates a discriminative module sharing the appearance encoder with the generative module. The high-quality cross-id composed images are fed back to the appearance encoder online, and used to improve the model for discriminative module. Zhang et al. [84] proposed the AlignedReID, which performs automatic part alignment during learning, without requiring extra supervision or pose estimation. Sun et al. [87] proposed a part-based convolutional baseline (PCB), consisting of several part-level features, with a refined part pooling to eliminate the outliers and further boost the performance in ReID. Wang et al. [114] deploy the same PCB as their visual feature streams, and estimate the spatial-temporal probability distribution stream as the constraint. The joint two stream approach reaches the state-of-the-art rank-1 accuracy of 98.1% on Market-1501 and 94.4% on DukeMTMC.

5.2.2 Adversarial Attacks

Adversarial Examples

Szegedy et al. [21] introduced the adversarial images, which can fool the Convolutional Neural Network (CNN)-based models, and cause misclassification by adding small perturbations to the original images. Goodfellow et al. [94] proposed fast gradient sign method (FGSM), which generates AEs in one step. Several works extended this by iteratively updating the AEs with multi-step attacks including the basic iterative method (BIM) [22], deep fool [95], momentum iterative method [23], Diverse Inputs Method (DIM) [96] and Translation-Invariant (TI) attacks [97]. Compared to FGSM, the iterative methods generate a smaller perturbation, which makes the adversarial examples even more imperceptible to human eye. Lu et al. [91] showed that AEs generated by Dispersion reduction (DR) are transferable, and can effectively attack different networks designed for different tasks.

Adversarial Attacks in Person ReID

Zheng et al. [90] proposed Opposite Direction Feature Attack (ODFA) to generate adversarial examples/queries for retrieval tasks such as person ReID. The idea is to push away the feature of the adversarial query in the opposite direction of the original feature. Wang et al. [115] proposed a learning-to-misrank formulation to perturb the ranking of the system output, which drops one of the best ReID performances from 91.8% to 1.4% after being attacked. Zheng et al. [116] extended the use of Dispersion Reduction (DR) to person ReID, and effectively attacked three different ReID networks using four different datasets.

5.2.3 Defense Frameworks

Papernot et al. [28] provide a comprehensive summary of work on the defense against adversarial examples. The defense work can be grouped into three broad categories: adversarial training, gradient masking and input transformation. Adversarial training introduces the discovered adversarial examples (AEs) and the corresponding ground truth labels into the training. Ideally, the model will learn how to differentiate the AEs from the ground truth and classify them into a new category. However, it suffers from the high cost of generating AEs, and at least doubles the training time due to the iterative retraining. Also, it assumes all the attacks are known and used when generating the AEs, which is not realistic for real-time scenarios. Gradient masking seeks to reduce the sensitivity of DNN models to small changes in inputs by forcing the DNN models to produce near-zero gradients [29]. Papernot et al. [28] concluded that methods designed to conceal gradient information are bound to have limited success because of the transferability of adversarial examples. Recent studies try to reduce the sensitivity to small changes in inputs by input transformation. Dimension reduction by using Principal Component Analysis(PCA) [30], image filtering by training an auto-encoder [31] and

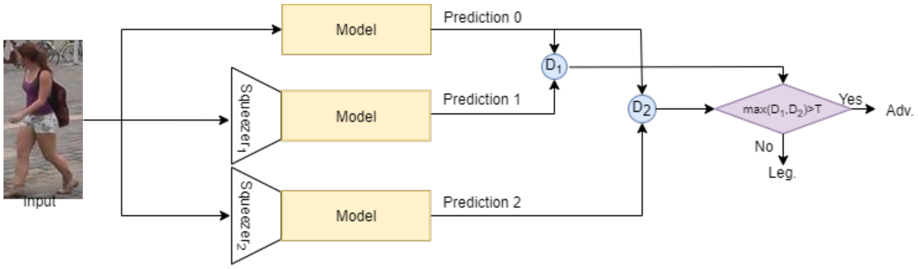


Fig. 5.1: The adversarial example detection based on feature squeezing

feature squeezing [32] are proposed to perform different transformations on the input data. However, much less attention has been paid to the defense of the ReID networks.

5.3 Proposed Method

In this work, we employ feature squeezing [32], together with a part-based convolutional baseline (PCB) [87] ReID model and combine two types of squeezing, (i) reducing the color depth of images and (ii) using non-local smoothing, to detect adversarial examples. The structure of the proposed approach is shown in Fig. 5.1. As will be discussed in Sec. 5.4, we have also experimented with a not parts-based ReID network (different from PCB) for the 'Model' seen in Fig. 5.1, and have shown the advantage of using the parts-based model for the detection of adversarial examples.

The detection mechanism here is based on the idea that the robustness of DNNs to local changes (e.g., squeezing, scale, position) does not generalize to the perturbations added by AEs, which have been empirically validated by previous AE detection works [31, 32, 117]. In other words, if a transformation is applied to an AE, the output of the network for the original AE and its transformed version will be very different, which is used to detect AEs as shown in Fig. 5.1.

5.3.1 Squeezing Color Bits

Most of the images in ReID datasets use 8-bits per color channel. However, it is difficult to tell the difference between the original images and the images using as few as 4 bits of color depth. Thus, we have performed 4,5,6 and 7-bit squeezing in our experiments. Since the original images are normalized, to reduce the color depth down to i-bit (where $4 \leq i \leq 7$), we only need to do pixel-wise multiplication by $2^i - 1$, round it, and then divide by $2^i - 1$ as shown in Eq. (5.1).

$$x^{sqz} = \left\lceil (x^{ori} * (2^i - 1)) \right\rceil / (2^i - 1), \tag{5.1}$$

where x^{sqz} is the squeezed image tensor, x^{ori} is the original image tensor, and i is the reduced bit depth $(4 \leq i \leq 7)$.

5.3.2 Squeezing Variations via Smoothing

Local smoothing by a median filter is particularly effective in removing sparsely-occurring black and white pixels in an image, while non-local smoothing is applied over a larger area of the image [32]. We apply a variant of a Gaussian filter as non-local smoothing.

5.3.3 Joint Detector

We perform n-bit squeezing and Gaussian smoothing as the two squeezers shown in Fig. 5.1. First, second and third rows of Fig. 5.2 show the body partition results of the PCB model, the partition results after 4-bit squeezing, and result of applying a 3x3 Gaussian smoothing to the segmented stripes, respectively.

The difference between the predictions based on the original image and the

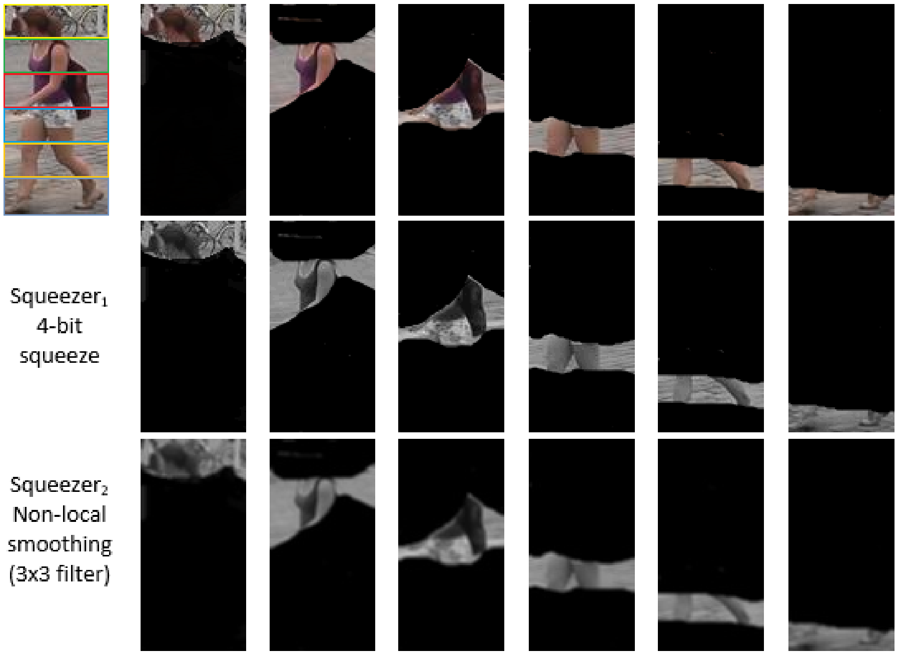


Fig. 5.2: Body parts segmented by PCB and their squeezed versions.

squeezed image is calculated as follows:

$$dist^{(\mathbf{x},\mathbf{x}_{sqz})} = ||s(\mathbf{x}) - s(\mathbf{x}_{sqz})||, \tag{5.2}$$

where $\mathbf{x}$ and $\mathbf{x}_{sqz}$ are the original input and the squeezed input, $s()$ is the softmax layer of the DNN model.

We can also combine multiple feature squeezers by the maximum distance as in Eq. 5.3.

$$dist^{joint} = max(dist^{(\mathbf{x},\mathbf{x}_{sqz1})}, dist^{(\mathbf{x}.\mathbf{x}_{sqz2})}, ...) \tag{5.3}$$

Similar to [32], we use the maximum distance assuming that different squeezers will be effective for different types of perturbations.

5.4 Experimental Results

In the training phase of the AE detector, an optimal threshold for $dist^{joint}$ is determined to detect AEs. It is obvious that a lower threshold value leads to a higher false positive rate which would be useless for many security sensitive tasks. We set a threshold value, which makes the false positive rate less than 5%, and report the true positive rate as the detection performance.

5.4.1 Datasets

We have employed three challenging and commonly used benchmark datasets to demonstrate the effectiveness of the proposed AE detection approach. These datasets are Market-1501 [88], CUHK03 [12], and DukeMTMCreID [112].

5.4.2 Victim Models

We first use two publicly available ReID models, PCB [87] and Open ReID [118], and attack them by the AEs generated by three different attack approaches. Open-ReID provides both Inception and Resnet50 as the backbone networks as well as some common loss functions. PCB model can take major backbones and output the convolutional features, with the help of refined pooling layer. It can provide a SOTA segmentation and boost the ReID performance.

We then use both PCB and OpenReID together with feature squeezing to evaluate their AE detection performance. We use ResNet50, as the backbone network, and triplet loss from Open-ReID library, since this combination achieves the best performance on all Market-1501, CUHK03, and DukeMTMCreID datasets [118].

Methods/mAP	Market-1501	CUHK03	DukeMTMC
PCB	81.6	57.5	69.2
OR	67.9	80.7	54.6
TI-DIM-PCB	22.6	19.3	16.6
TI-DIM-OR	24.8	20.6	13.9
DR-PCB	15.4	8.1	10.8
DR-OR	11.7	10.3	9.5
MR-PCB	4.3	2.1	1.2
MR-OR	4.4	0.9	1.5

Table 5.1: The mAP values for two ReID networks before and after being attacked by three attack methods. Results show the effectiveness of the attacks in degrading ReID performance. The lower mAP value means better attack performance.

5.4.3 Results and Discussion

Tab. 5.1 shows the results of attacking two different victim models by AEs generated by TI-DIM [97], Dispersion Reduction(DR) [91] and Mis-Ranking(MR) [115]. For all the attacks, the perturbation threshold is set to be $\epsilon = 16$ for consistency.

The first two rows of Tab. 5.1 show the baseline mAP for the victim models on three different datasets. The mAP values for PCB are 81.6%, 57.5% and 69.2%, and the mAP values for OpenReID (OR) are 67.9%, 80.7% and 54.6% for Market1501, CUHK03 and DukeMTMC datasets, respectively. Rows 3-4 are the mAP after the victim models are attacked by TI-DIM. The mAP values for PCB drop to 22.6%, 19.3% and 16.6%, while mAP values for OR drop to 24.8%, 20.6% and 13.9% for Market1501, CUHK03 and DukeMTMC-ReID datasets, respectively. Similarly, Row 5-8 are mAP values ater the victim models are attacked with the DR and MR attacks. As can be seen, all of the attacks degrade the performance of person ReID significantly.

Table 5.2 shows the AE detection accuracy when two different ReID models (one part-based, one not part-based) are used together with a combination of two kinds of feature squeezing. 100 AEs are generated by each of the TI-DIM, DR and

Accuracy	ReID Model		Market1501	CUHK03	DukeMTMC
TI-DIM	OpenReID	4-bit	69.1	55.2	56.8
		5-bit	68.7	55.1	56.6
		6-bit	68.5	56.1	55.3
		7-bit	69.2	55.8	55.7
	PCB	4-bit	79.5	70.4	71.7
		5-bit	80.1	69.9	72.0
		6-bit	80.1	**70.6**	**72.1**
		7-bit	**80.3**	68.9	71.8
DR	OpenReID	4-bit	68.2	56.9	57.4
		5-bit	68.1	56.8	57.5
		6-bit	66.9	55.4	57.4
		7-bit	69.7	56.8	57.6
	PCB	4-bit	80.2	68.9	77.6
		5-bit	80.5	70.2	74.1
		6-bit	**81.2**	68.7	74.6
		7-bit	79.7	**71.3**	**77.8**
MR	OpenReID	4-bit	65.6	57.1	54.8
		5-bit	64.9	57.5	53.7
		6-bit	66.9	56.8	54.4
		7-bit	65.1	56.9	55.0
	PCB	4-bit	**80.7**	69.8	69.1
		5-bit	79.1	64.2	69.0
		6-bit	79.9	**70.2**	72.6
		7-bit	80.0	65.5	**73.9**

Table 5.2: Accuracy of detecting adversarial examples generated by TI-DIM, DR and MR on different datasets when two different ReID models are used. The best results are shown in bold. n-bit refers to n-bit feature squeezing and Gaussian smoothing being used as two squeezers.

MR from each of the three different datasets. As shown in [32], combining bit-wise squeezing and smoothing outperforms bit-wise squeezing alone. So, in our experiments, we used 4-7 bit-wise squeezing together with the Gaussian smoothing.

As seen in Tab. 5.2, using the part-based PCB model together with feature squeezing provides much better performance for AE detection compared to using Open ReID, which is not part-based. For instance, for the TI-DIM attack, the best AE detection accuracies with OpenReID model are 69.2%, 56.1% and 56.8% on

Market-1501, CUHK03 and DukeMTMC datasets, respectively. When we use the part-based PCB model, the detection accuracy improves to 80.3, 70.6 and 72.1 on Market-1501, CUHK03 and DukeMTMC datasets, respectively. A similar trend is observed for the DR and MR attacks as well. Part-based PCB model increases the best detection accuracies by 11.5%, 14.4%, 20.2% for the DR attack, and by 13.8%, 12.7%, 18.9% for the MR attack on three datasets.

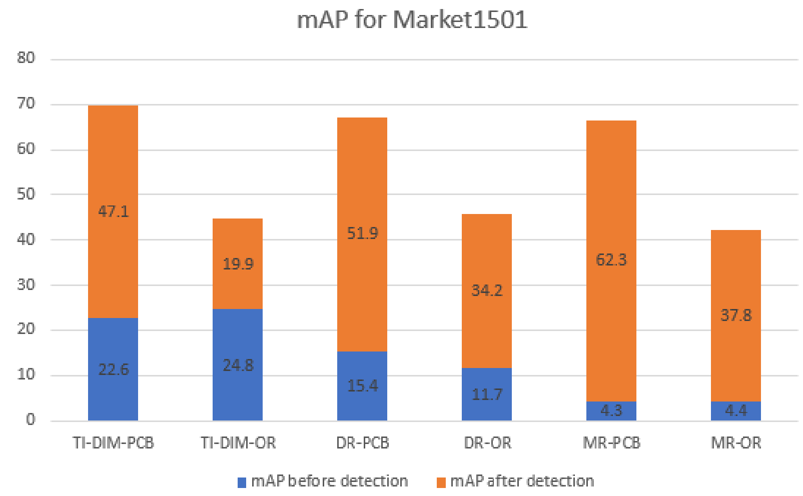


Fig. 5.3: The mAP values after AE detection on the Market1501 dataset. The blue bars are the mAP after each attack, and the orange bars are the improvement after the AE detection.

To further demonstrate the effectiveness of AE detection and its benefit on the performance of ReID networks, we have performed an experiment similar to [32], but for person ReID. More specifically, we generated 100 adversarial examples by each of the three attacks in Tab. 5.2. We used these 300 AEs together with 100 original/benign input images as the probe images feeding them into a PCB and then Open-ReID network. The results are shown in Fig. 5.3. As can be seen, with the PCB model, the mAP after AE detection can reach close to 70%, while for Open-ReID, the mAP can only improve to around 46%.

5.5 Conclusion

In this chapter, we have presented a novel defense method that utilizes parts-based feature squeezing to detect adversarial examples presented to ReID networks. We have applied two types of squeezing together to segmented body parts to better detect adversarial examples. We have compared the detection performance of the proposed body part-based approach with a ReID method that is not parts-based. We have also evaluated the mAP before and after detecting adversarial examples. Experimental results have shown that the proposed method can effectively detect the adversarial examples, and has the potential to avoid significant decreases in person ReID performance caused by adversarial attacks.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Given a query image or video of a person of interest, person re-identification is the process of identifying the same person in images or videos captured by a different camera, or by the same camera at a different time. Person re-identification is very important to be able to continuously track a person of interest across multiple cameras with disjoint views. In this dissertation, we have first presented a novel approach to better model the appearance variation across disjoint camera views, and improve the performance of any person re-identification approach that incorporates color/brightness histograms and appearance models. We have proposed building a codebook of brightness transfer functions (BTF), and also an ordering and trimming criteria to increase computational efficiency. We have performed an extensive set of experiments on different commonly used datasets. Results have shown that the proposed method outperforms other BTF-based approaches. Moreover, the proposed approach was incorporated into four different state-of-the-art person re-identification methods, and an increased top rank matching rate has been obtained for all methods and on all datasets, supporting our initial argument above.

Then, we have extended the person re-identification from across a set of static cameras to a larger scale via crowdsourcing and using the images or videos captured by people's cellphones. If it is possible to have access to large number of images/videos captured at different locations over time, this would potentially allow following a person over a much larger area for extended periods of time. We have proposed a method, which combines R-CNN based person detection with the GPU implementation of color histogram and SURF-based re-identification. The GPS location in the image/video EXIF information has been used to obtain a spatio-temporal model for the path taken by the target, and these locations have been displayed on a map. All the processing was performed on a GPU.

With the significant advances in deep neural networks (DNN), and their success in image classification and object detection tasks, researchers have proposed DNN-based approaches for the person re-identification problem, and achieved state-of-the-art performance. However, it has also been showed that neural networks are vulnerable to carefully crafted adversarial examples, and can be easily deceived. We have presented a black-box and effective attack model, which is based on dispersion reduction, and does not rely on task-specific loss functions and label queries. We have used the adversarial examples generated by this approach to attack two different state-of-the-art person Re-ID models to demonstrate the vulnerability of multiple state-of-the-art person re-identification approaches to this attack, and draw attention to the existing security risks. We have also compared the performance of our attack approach with two other state-of-the-art attack models. The results demonstrate the effectiveness of the proposed dispersion reduction attack on two state-of-the-art person re-identification models. It has outperformed other state-of-the-art attack models by a large margin, and resulted in the most drop in the mean average precision values.

After presenting an effective attack method, and showing the vulnerability of the state-of-the-art person re-identification approaches, we have focused on the defense, and developed a method to detect the adversarial examples presented to a person re-identification network. This method is based on parts-based feature squeezing. We have applied two types of squeezing to segmented body parts to better detect adversarial examples. We have compared the detection performance of our proposed body part-based approach with a re-identification method that is not parts-based. We have also evaluated the mean average precision before and after detecting adversarial examples. Experimental results have shown that the proposed method can effectively detect the adversarial examples, and has the potential to avoid significant decreases in person re-identification performance caused by adversarial attacks.

While we presented an effective attack method, as well as a defense approach to detect the adversarial examples, there are other areas that can be investigated as future work. The robustness of neural network-based re-identification methods can be further explored. Cross-dataset person-identification refers to training a model on one dataset and then testing it on different datasets. Cross-dataset transferability of the dispersion reduction attack, and its detection by the cross-dataset person re-identification models can be further investigated. As for the crowdsourcing application, more features can be explored to improve the re-identification performance.

REFERENCES

[1] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PET)*, Rio de Janeiro, Brazil, Oct. 14. 2007.

[2] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," in *International Conference on Pattern Recognition (ICPR)*, Tsukuba, Japan, Nov. 11. 2012, pp. 2367–2370.

[3] X. Wang and R. Zhao, *Person Re-identification: System Design and Evaluation Overview*, pp. 351–370, Springer London, London, 2014.

[4] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, US., Jun. 16. 2012, pp. 2666–2672.

[5] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 3, pp. 653–668, March 2013.

[6] C. Liu, C. C. Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *2013 IEEE International Conference on Computer Vision (CVPR)*, Portland, Oregon, US., Jun. 25. 2013, pp. 441–448.

[7] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon,US., Jun. 25. 2013, pp. 3610–3617.

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proceedings of the British Machine Vision Conference (BMVC)*, University of Dundee, UK., Aug. 29. 2011.

[9] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon, US., Jun. 25. 2013.

[10] B. Prosser, W. Zheng, G. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proceedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, UK., Aug. 31. 2010, pp. 21.1–21.11.

[11] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proceedings of European Conference on Computer Vision (ECCV)*, Florence, Italy, Oct. 7. 2012, pp. 402–412.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, US., Jun. 24. 2014, pp. 152–159.

[13] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, US., Jun. 24. 2014, pp. 144–151.

[14] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), Boston, Massachusetts, US., Jun. 8. 2015, pp. 3707–3715.

[15] F. Porikli, "Inter-camera color calibration by correlation model function," in *Proceedings of International Conference on Image Processing (ICIP)*, Barcelona, Catalonia, Spain, Sep. 14. 2003, vol. 2, pp. II–133–6.

[16] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, USA, Jun. 20. 2005, vol. 2, pp. 26–33.

[17] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proceedings of the British Machine Vision Conference (BMVC)*, Leeds, UK., Sep. 1. 2008.

[18] A. Bhuiyan, B. Mirmahboub, A. Perina, and V. Murino, "Person re-identification using robust brightness transfer functions based on multiple detections," in *Proceedings of Image Analysis and Processing (ICIAP)*, Genoa, Italy, Sep. 7. 2015, pp. 449–459.

[19] A. Gala and S. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, pp. 270–286, Apr. 2014.

[20] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, San Jose, California, US., May. 23. 2016, pp. 39–57.

[21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, Apr. 14. 2014.

[22] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *International Conference on Learning Representations (ICLR) Workshop*, Banff, AB, Canada, Apr. 14. 2017.

[23] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, US., Jun. 18. 2018, pp. 9185–9193.

[24] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, “Adversarial examples for semantic segmentation and object detection,” in *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct. 22. 2017, pp. 1378–1387.

[25] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, and T. Kohno, “Physical adversarial examples for object detectors,” in *USENIX Workshop on Offensive Technologies (WOOT)*, Baltimore, MD, Aug. 15. 2018.

[26] X. Wei, S. Liang, N. Chen, and X. Cao, “Transferable adversarial attacks for image and video object detection,” in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, Aug. 10. 2019, pp. 954–960.

[27] A. Arnab, O. Miksik, and P. Torr, “On the robustness of semantic segmentation models to adversarial attacks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, US., Jun. 18. 2018.

[28] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “Sok: Security and privacy in machine learning,” in *IEEE European Symposium on Security and Privacy (EuroSP)*, London, UK., Apr. 24. 2018, pp. 399–414.

[29] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” *arXiv preprint arXiv:1412.5068*, 2014.

[30] A. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in *conference on Information Sciences and Systems*, Princeton, US., May. 21. 2018, pp. 1–5.

[31] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of ACM SIGSAC conference on Computer and Communications Security*, Dallas, US., Oct. 30. 2017, pp. 135–147.

[32] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[33] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, Colorado, US., Jun. 20. 2011, pp. 649–656.

[34] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Asian Conference on Computer Vision (ACCV)*, Queenstown, New Zealand, Nov. 8. 2011, pp. 501–512.

[35] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision (ACCV)*, Daejeon, Korea, Nov. 5. 2013, pp. 31–44.

[36] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, California, US., Jun. 13. 2010, pp. 2360–2367.

[37] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *Proceedings of the British Machine Vision Conference (BMVC)*, Guildford, UK., Sep. 3. 2012, p. 11.

[38] C. Liu, S. Gong, C. Loy, and X. Lin, “Person re-identification: What features are important?,” in *Proceedings of European Conference on Computer Vision (ECCV)*, Florence, Italy, Oct. 7. 2012, pp. 391–401.

[39] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Proceedings of European Conference on Computer Vision (ECCV)*, Marseille, France, Oct. 12. 2008, pp. 262–275.

[40] R. Layne, T. Hospedales, and S. Gong, “Person re-identification by attributes,” in *Proceedings of the British Machine Vision Conference (BMVC)*, Surrey, UK., Sep. 3. 2012, pp. 24.1–24.11.

[41] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, “Attribute-restricted latent topic model for person re-identification,” *Pattern Recognition*, vol. 45, no. 12, pp. 4204–4213, Dec. 2012.

[42] E. D. Cheng and M. Piccardi, “Matching of objects moving across disjoint cameras,” in *International Conference on Image Processing (ICIP)*, Atlanta, Georgia, US., Oct. 8. 2006, pp. 1769–1772.

[43] C. Madden and M. Piccardi, “A framework for track matching across disjoint cameras using robust shape and appearance features,” in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, London, UK., Sep. 5. 2007, pp. 188–193.

[44] H.U. Chae and K.H. Jo, “Appearance feature based human correspondence under non-overlapping views,” in *International Conference on Intelligent Computing (ICIC)*, Ulsan, South Korea, Sep. 16. 2009, pp. 635–644.

[45] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *IEEE Conference on Com-*

puter Vision and Pattern Recognition (CVPR), Boston, Massachusetts, US., Jun. 8. 2015, pp. 2197–2206.

[46] Y. C. Chen, X. Zhu, W. S. Zheng, and J. H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. PP, no. 99, pp. 1–1, Feb. 2017.

[47] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, US., Jun. 26. 2016, pp. 1363–1372.

[48] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, US., Jun. 16. 2012, pp. 2288–2295.

[49] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[50] J. Iivarinen and A. Visa, "Shape recognition of irregular objects," in *Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling*, 1996, pp. 25–32.

[51] J. Jain, S. Sahoo, S. R. Mahadeva Prasanna, and G. Siva Reddy, "Modified chain code histogram feature for handwritten character recognition," in *Advances in Computer Science and Information Technology. Networks and Communications (CCSIT)*, Bangalore, India, Jan. 2. 2012, pp. 611–619.

[52] W. Li and X. Wang, "Locally aligned feature transforms across views," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon, US., Jun. 25. 2013.

[53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, Sept 2010.

[54] Zheng Wang, “Re-id resources,” https://wangzwhu.github.io/home/re_id_resources.html.

[55] P. Viola and M. Jones, “Robust real-time face detection,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Vancouver, British Columbia, Canada, Jul. 7. 2001, vol. 2, pp. 747–747.

[56] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, US., Jun. 20. 2005, vol. 1, pp. 886–893.

[57] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK., Sep. 7. 2009.

[58] X. Wang, T. Han, and S. Yan, “An hog-lbp human detector with partial occlusion handling,” in *IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, Sep. 29. 2009, pp. 32–39.

[59] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, “Multiple kernels for object detection,” in *IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, Sep. 29. 2009, pp. 606–613.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 60, no. 6, pp. 84–90, May 2012.

[61] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), Columbus, Ohio, US., Jun. 24. 2014, pp. 580–587.

[62] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, vol. 28.

[63] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, Aug. 23. 2010, pp. 1413–1416.

[64] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[65] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, Oct. 14. 2007, pp. 1–8.

[66] N. Gheissari, T.B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, New York, US., Jun. 17. 2006, vol. 2, pp. 1528–1535.

[67] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013.

[68] N. Martinel, G. Foresti, and C. Micheloni, "Person reidentification in a distributed camera network framework," *IEEE Transactions on Cybernetics*, vol. 47, no. 11, pp. 3530–3541, 2017.

[69] N. Martinel and C. Micheloni, "Classification of local eigen-dissimilarities for person re-identification," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 455–459, 2015.

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

[71] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 5, pp. 1095–1108, 2015.

[72] Z. Akhtar, N. Martinel, C. Micheloni, and G. Foresti, "Mobile re-identification based on local features analysis," in *Proceedings of the International Conference on Distributed Smart Cameras (ICDSC)*, Venezia Mestre, Italy, Nov. 4. 2014.

[73] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury, "Re-identification in the function space of feature warps," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 8, pp. 1656–1669, 2015.

[74] G. Lisanti, I. Masi, A. Bagdanov, and A. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 8, pp. 1629–1642, 2015.

[75] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, Boston, Massachusetts, US., Jun. 8. 2015, pp. 33–40.

[76] S. Bąk, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *Proceedings of Eu-*

ropean Conference on Computer Vision (ECCV), Florence, Italy, Oct. 7. 2012, pp. 806–820.

[77] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proceedings of European Conference on Computer Vision (ECCV)*, Florence, Italy, Oct. 7. 2012, pp. 780–793.

[78] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of ACM international conference on Multimedia*, 2014, pp. 675–678.

[79] X. Chen, P. Wei, W. Ke, Q. Ye, and J. Jiao, "Pedestrian detection with deep convolutional neural network," in *Asian Conference on Computer Vision Workshops (ACCV)*, Singapore, Nov. 1. 2014, pp. 354–365.

[80] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision Image Understanding*, vol. 110, no. 3, pp. 346–359, June 2008.

[81] P. Harvey, "Exiftool by phil harvey," `http://www.sno.phy.queensu.ca/~phil/exiftool/`, 2008.

[82] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, US., Jun. 16. 2019.

[83] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," *Pattern Recognition*, vol. 94, pp. 53–61, 2019.

[84] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[85] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *arXiv preprint arXiv:1711.10658*, 2017.

[86] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, 2017.

[87] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 8. 2018, pp. 501–518.

[88] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 11. 2015.

[89] S. Bai, Y. Li, Y. Zhou, Q. Li, and P. Torr, "Adversarial metric attack and defense for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Oct. 2020.

[90] Z. Zheng, L. Zheng, Z. Hu, and Y. Yang, "Open set adversarial examples," arXiv preprint, arXiv 1809.02681, 2018.

[91] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, and S. Velipasalar, "Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, US., Jun. 16. 2020, pp. 940–949.

[92] B. Xie, X. Wu, S. Zhang, S. Zhao, and M. Li, "Learning diverse features with part-level resolution for person re-identification," *ArXiv*, vol. abs/2001.07442, 2020.

[93] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct. 27. 2019.

[94] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[95] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, US., Jun. 26. 2016, pp. 2574–2582.

[96] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2730–2739.

[97] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, US., Jun. 16. 2019.

[98] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of ACM on Asia conference on Computer and Communications Security (AsiaCCS)*, Abu Dhabi, United Arab Emirates, Apr. 2. 2017, ACM, pp. 506–519.

[99] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[100] J. Uesato, B. O'Donoghue, A. Oord, and P. Kohli, "Adversarial risk and the dangers of evaluating against weak attacks," *arXiv preprint arXiv:1802.05666*, 2018.

[101] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," *arXiv preprint arXiv:1804.08598*, 2018.

[102] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248*, 2017.

[103] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[104] Carroll C. and Paul T., "Nist/sematech e-handbook of statistical methods," http://www.itl.nist.gov/div898/handbook/, 2012.

[105] Erbloo, "Dispersion reduction," `https://github.com/erbloo/Dispersion_reduction`.

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, US., Jun. 26. 2016, pp. 770–778.

[107] "Reid leaderboard," `https://github.com/handong1587/handong1587.github.io/blob/master/_posts/deep_learning/2015-10-09-re-id.md`.

[108] layumi, "Dg-net," `https://github.com/NVlabs/DG-Net`.

[109] michuanhaohao, "Alignedreid," `https://github.com/michuanhaohao/AlignedReID`.

[110] AI-NERC-NUPT, "Plr-osnet," `https://github.com/AI-NERC-NUPT/PLR-OSNet`.

[111] dongyp13, "Translation-invariant-attacks," https://github.com/dongyp13/Translation-Invariant-Attacks.

[112] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking (ECCV)*, Amsterdam, The Netherlands, Oct. 8. 2016.

[113] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, US., Jun. 18. 2018, pp. 79–88.

[114] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," *Proceeding of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8933–8940, Jan. 27. 2019.

[115] H. Wang, G. Wang, Y. Li, D. Zhang, and L. Lin, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, US., Jun. 16. 2020, pp. 339–348.

[116] Y. Zheng, Y. Lu, and S. Velipasalar, "An effective adversarial attack on person re-identification in video surveillance via dispersion reduction," *IEEE Access*, vol. 8, pp. 183891–183902, 2020.

[117] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao, "Foveation-based mechanisms alleviate adversarial examples," *arXiv preprint arXiv:1704.01155*, 2016.

[118] T. Xiao, "Open-reid," https://github.com/Cysu/open-reid.

VITA

NAME OF AUTHOR: Yu Zheng

Yu Zheng (S'12) received the M.S. degree in electrical engineering from Syracuse University, Syracuse, NY, USA in 2012, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include object detection/tracking, multi-camera systems, bio-medical image processing, and computer vision.