國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

對光線變化具有強健適應的人物重新識別系統

輔以基於群聚的損失函數

Person Re-Identification Robust to Illumination Change with Clustering-based Loss Function

劉宇閎

Yu-Hung Liu

指導教授：傅立成 博士

Advisor: Li-Chen Fu, Ph.D.

中華民國108年8月

August, 2019

1. 誌謝

回首過去的自己，碩班的這兩年來，有了太多的改變。從一開始對於影像處理沒有任何基礎認知到最後完成了一個研究。在這過程中很感謝家人、女朋友還有還有朋友的支持。

再來也感謝我的指導教授 傅立成教授，在這兩年來帶給我對於學術研究精神和態度的正確指引。老師致力於把學術研究的成果導向日常生活應用的精神也讓我學習到了很多實作的技巧以及當遇到問題時解決的應變能力，謝謝老師。

此外也要感謝在百忙中撥空來參與我的口試考試的口試委員：傅楸善教授、張文中教授、黃正民教授以及王鈺強教授，在口試過程中對於研究給的各種意見以及如何把碩士論文詮釋得更完整給了相當多的意見，這些意見也讓也讓我也讓我的碩士論文更加完整。也感謝資工影像組碩二的夥伴們：禹齊、子翔還有恩德，在研究的過程中我們遇到了許多挫折以及困境，大家一起互相給意見、彼此分享技術；也謝謝電機ACL其他夥伴，這兩年大家共同經歷了許多的展示以及會議，大家一起學習、成長 還有玩樂。

另外也謝謝安陞學長在這兩年給了我許多的協助還有指導，幫影像組減輕了不少的負擔，讓我們可以減輕許多的壓力把更多的精力專注在研究上。也謝謝影像組的學弟郭權、睿庭、靈風、炳彰還有Tommy幫我處理了很多的雜事還有協助我論文的校稿，在我許多需要幫忙的時候都二話不說馬上答應我，在未來的研究路上還會有很多的挑戰，大家可以在這過程中得到許多經驗也相信各位可以在未來運用這些經驗來解決研究上的問題。

宇閎 August 8, 2019

1. 摘要

近年來，人物重新識別系統受到大量的關注，因為其有廣大的應用場域像是智慧家庭、健康照護以及監視系統。但是隨著視角的改變以及拍攝相機的位置不同，人的輪廓外觀也會跟著不同，這造成了從不同的視角進行行人追蹤仍然是個挑戰。

除此之外，在實際應用的場域中在各個相機之間的光線照射位置及程度都是不同的，而在當前的人物重新識別模組中往往只會透過現有的資料集來學習這使得模型不具有能應付光亮變化的能力。最後我們在分析了當前最先進的人物重新識別期刊文章後我們也發現，目前大多數的研究為了要訂出閾值來區分正樣本及負樣本，都會採用指標損失函數來做模型優化，雖然指標損失函數可以有效的區分正負樣本，但使用這種損失函數卻需要面臨因時間複雜度過高而使得訓練冗長，導致我們將模型移轉到新環境時即使已經蒐集好新環境的資料仍然需要花費大量時間進行模型校正。

首先若要解決光亮變化的問題，最直觀的方式就是搜集大量光亮變化的人物重新識別資料集，但是這件事情卻是相當困難且需耗費相當多時間以及人力的。基於上述原因，本論文提出了一種透過合成資料來協助模型訓練以提取無關光亮變化的特徵向量。而針對指標損失函數時間複雜度過高的缺點，本研究也提出了一種基於群聚的損失函數以降低時間複雜度並且效能更優於指標損失函數。。並在最終實驗也證明本論文提出的方法及損失函數在人物重新識別的任務中超過其他行人重新識別方法。

**關鍵字：**深度學習、資料檢索、人物重新識別、聚合損失函數

1. ABSTRACT

Nowadays, person re-identification has raised lots of attention in the area of computer vision, because of its wide applications, including smart home, elderly care, and surveillance systems. From different viewpoints, the shape of the human body looks completely different, hence tracking human from different camera remains a challenging problem.

In addition, the locations and levels of light illumination can be different among cameras in the field of actual application. However, the existing person re-identification module is often learned only through the available dataset, which make model fail to be robust to situations with illumination change. Finally, after analyzing the recent literature on pedestrian re-recognition, we also found that most of the current researches use the metric loss function to optimize the model with an appropriate threshold to distinguish the positive samples from the negative ones. Despite the metric loss function can perform objective distinction as mentioned, it remains to have a disadvantage of having highly complex, which make the training process lengthy.

First of all, to solve the problem of brightness changes, the most intuitive way is to collect an even larger person re-identification dataset subject to various brightness levels, which however is very expensive to collect and label. Therefore, this thesis proposes an illumination-invariant feature vector that assists model training based on synthetic data. To remove the shortcomings of the time complexity of the metric loss function, we propose the clustering-based loss function to reduce the time complexity, and we also show that the performance of the proposed loss function is better than metric loss function. In the final experiment, it is also proved that the proposed method in this thesis excels the state-of-the-art methods on resolving person re-identification problems.

**Keywords:** Deep learning, Information retrieval, Person re-identification, Clustering-based loss function

1. TABLE OF CONTENTS

口試委員會審定書 #

[誌謝 I](#_Toc15689241)

[摘要 II](#_Toc15689242)

[ABSTRACT III](#_Toc15689243)

[TABLE OF CONTENTS V](#_Toc15689244)

[LIST OF FIGURES VIII](#_Toc15689245)

[LIST OF TABLES XI](#_Toc15689246)

[Chapter 1 Introduction 1](#_Toc15689247)

[1.1 Motivation 1](#_Toc15689248)

[1.2 Literature Review 3](#_Toc15689249)

[1.2.1 Human Detection 3](#_Toc15689250)

[1.2.2 Person Re-Identification 7](#_Toc15689251)

[1.2.3 Domain Adaptation 10](#_Toc15689252)

[1.3 Contributions 11](#_Toc15689253)

[1.4 Thesis Organization 12](#_Toc15689254)

[Chapter 2 Preliminaries 14](#_Toc15689255)

[2.1 Cluster Analysis and K-means 14](#_Toc15689256)

[2.2 Convolutional Neutral Network 17](#_Toc15689257)

[2.2.1 Convolutional Layers 18](#_Toc15689258)

[2.2.2 Residual Network 20](#_Toc15689259)

[2.3 Real-time Pose Estimation Module 23](#_Toc15689260)

[2.4 Information Retrieval 24](#_Toc15689261)

[Chapter 3 Person Re-Identification 26](#_Toc15689262)

[3.1 Learn an Illumination-Invariant Feature 26](#_Toc15689263)

[3.1.1 Synthetic Dataset 26](#_Toc15689264)

[3.1.2 Learn from synthetic data 28](#_Toc15689265)

[3.1.3 Domain Adaptation by Adversarial Learning 30](#_Toc15689266)

[3.2 Assist by Clustering 33](#_Toc15689267)

[3.2.1 Clustering Loss 34](#_Toc15689268)

[3.2.2 Adaptive Weighted Clustering Loss 36](#_Toc15689269)

[3.2.3 Hard Clustering Mining 37](#_Toc15689270)

[Chapter 4 ACL Re-Identification Dataset 39](#_Toc15689271)

[4.1 Environment setting 39](#_Toc15689272)

[4.2 Preprocessing 42](#_Toc15689273)

[Chapter 5 Experiments 46](#_Toc15689274)

[5.1 Configuration 46](#_Toc15689275)

[5.2 Implementation Details 47](#_Toc15689276)

[5.2.1 Network design 47](#_Toc15689277)

[5.2.2 Training Details 48](#_Toc15689278)

[5.3 Person Re-Identification Dataset 50](#_Toc15689279)

[5.3.1 Market-1501 Dataset 50](#_Toc15689280)

[5.3.2 DukeMTMC-reID Dataset 51](#_Toc15689281)

[5.3.3 Evaluation Metrics 53](#_Toc15689282)

[5.4 Cross-Illumination Classification Result 54](#_Toc15689283)

[5.5 Person Re-Identification Result 55](#_Toc15689284)

[5.5.1 Ablation study 55](#_Toc15689285)

[5.5.2 The Result of Market-1501 Dataset 58](#_Toc15689286)

[5.5.3 The Result of DukeMTMC-reID Dataset 60](#_Toc15689287)

[5.5.4 The Result of ACL-reID 61](#_Toc15689288)

[Chapter 6 Conclusion and Future Works 62](#_Toc15689289)

[REFERENCE 63](#_Toc15689290)

1. LIST OF FIGURES

Figure 1‑1 Each row means the same human but taken under from different camera. 2

Figure 1‑2 The background subtraction result [7] 3

Figure 1‑3 The network architecture of RCNN [10] 4

Figure 1‑4 The difference between YOLO-v1 and YOLO-v2 5

Figure 1‑5 The architecture of YOLO-v1 [13]. 6

Figure 1‑6 The multi-channel part-based convolutional networks architecture [16] 7

Figure 1‑7 The network design for Pose-aware Multi-shot Matching [21]. 8

Figure 1‑8 The different between traditional triplet loss and adaptive weights triplet loss [25]. 9

Figure 1‑9 Domain adaptation using gradient reversal layer [31]. 10

Figure 2‑1 Clustering 2D data points from 4 clusters 14

Figure 2‑2 The data point assignment in first iteration. 15

Figure 2‑3 The k-means process. (a) means the given randomly initial center point, (b) represent the clustering results obtained by (a), we can re-calculate the new center points based-on clustering result after get (b) which is represented by (c) and so on. 16

Figure 2‑4 A standard classification task CNN architecture comprised of several convolutional layers and fully connected layers, the end-to-end training strategy can guarantee that CNN and fully connected layer are optimized simultaneously. 18

Figure 2‑5 The standard convolution operation with stride equal to 1 19

Figure 2‑6 The network design for AlexNet and Standard VGG16. 20

Figure 2‑7 The difference between traditional convolution block and residual block. 21

Figure 2‑8 The family of ResNet [46], each layer contains multi residual block. The model is named according to the number of layers it has. For example, we named 50-layer as ResNet-50. 22

Figure 2‑9 The architecture of the Openpose [26] 22

Figure 2‑10 Pose estimation process for Openpose[26] 23

Figure 2‑11 The idea of the information retrieval on Person Re-identification 24

Figure 3‑1 The image capture from SyRI dataset. 27

Figure 3‑2 Architecture for learning appearance feature. 28

Figure 3‑3 Visualization of illumination-invariant feature extracted from SyRI dataset. Each cluster represents the features with same identity from different lighting. Due to the limitation of the palette, same color may appear multiple times. 29

Figure 3‑4 Domain shift visualization which is performing dimensionality reduction by using t-SNE. Orange points is illumination invariant feature extracted from real data, blue points are illumination invariant feature extracted from synthetic data. 30

Figure 3‑5 The complete architecture for learning illumination-invariant feature with domain adaptation. The blue line means the flow for synthetic data and the red line is for real data, respectively. 31

Figure 3‑6 The flow of the proposed method. 34

Figure 3‑7 The idea of clustering loss’s variance term (a) and distance term (b) 35

Figure 3‑8 The different between uniform weighted and softmax weighted, the length of the arrow represents force. 36

Figure 3‑9 The difference between origin batch construction and HCM 38

Figure 4‑1 The camera topology in CSIE-Der Tian Hell’s second floor 39

Figure 4‑2 The example of each cameras 40

Figure 4‑3 The Example of different pose and different human orientation 41

Figure 4‑4 Pipeline of dataset preprocessing 43

Figure 4‑5 The difference between without padding (b) and with padding (c). 43

Figure 4‑6 The junk image sample, blue box represents the Openpose detection result, the red box denotes the cropped region which is padding from the blue box. 44

Figure 4‑7 The distribution of the ACL-reID data 45

Figure 5‑1 The hold architecture of the Encoder 47

Figure 5‑2 The training stages on our proposed method 49

Figure 5‑3 The different identities collected from different cameras. 50

Figure 5‑4 Sample images of the distractor dataset (junk) [61] 51

Figure 5‑5 The camera position setting for DukeMTMC-reID [27] 52

Figure 5‑6 The image distribution on DukeMTMC-reID training set. 52

Figure 5‑7 The Cumulative Match Characteristic (CMC) curve. 53

1. LIST OF TABLES

Table 2‑1 Confusion matrix 25

Table 5‑1 Specification of our experiment environment 46

Table 5‑2 Identity classification accuracy on SyRI dataset. 54

Table 5‑3 Comparison of different kinds of loss and the proposed loss function on the Market-1501 [61] and DukeMTMC-reID [62]. \* denotes the data of the experiment results are copied from the paper. 55

Table 5‑4 Ablation studies of IIF. DA is acronym for domain adaptation. 56

Table 5‑5 Ablation studies of hard clustering mining 57

Table 5‑6 Ablation studies of the proposed method 58

Table 5‑7 Comparison of the state-of-the-art methods on the market-1501 59

Table 5‑8 Comparison of the state-of-the-art methods on the DukeMTMC-reID 60

Table 5‑9 Comparison of the person re-identification on ACL-reID dataset 61

# Introduction

In this chapter, we first describe the motivation of this work in Section 1.1. A complete literature review is presented in Section 1.2. In Section 1.3, we highlight our contributions, and the organization of this thesis is presented in Section 1.4.

## Motivation

Recently, person re-identification(Re-ID) [1] has raised lots of attention in the area of computer vision, due to its wide applications, such as elderly care, smart home, and surveillance systems. The objective of person re-identification is given a person-of-interest (query) image, and Re-ID is intended to retrieve person images from multiple cameras that correspond to the same person with different view perspectives. Over the last decade, to extract the features of human images, there are some well-known approaches that utilize hand-crafted features, such as Histogram of Oriented Gradient (HOG) [2] and Scale-Invariant Feature Transform (SIFT) [3] feature. However, these kind of methods are hard to handle cross-camera person re-identification problem, and their performances drop sharply when the camera’s viewpoint is changed. This is because hand-craft features are view dependent but the appearance of a person will change when he/she moves to the fields of different cameras and is seen from different perspectives as shown in Figure 1‑1.

It has become a general fact that, deep learning methods can obtain more effective features in different tasks since the learned models can extract high-level representative feature from data through training instead of the hand-crafted features. Since 2010, Felzenszwalb *et al*. [4] proposed a novel method combining the HOG [2] feature and Support Vector Machine (SVM) [5], and thus machine learning methodology go beyond the traditional hand-crafted feature method in VOC competition [6]. Due to the development of powerful learning-based human detector, most researches design the Re-ID network which only focuses on appearance feature extraction.

|  |
| --- |
|  |
| Figure 1‑1 Each row means the same human but taken under from different camera. |

In our prior knowledge, face recognition has already been developed. However, person re-identification technology is still hard to apply to real world system due to the high computational cost which is caused by the deep model, and the feature extracted by the model cannot handle the lighting variance. Another problem is the time complexity of fine-tuning the pre-trained model. Although we can collect the new environment person re-identification to fine-tune our model from the training domain to the application domain, the computation of the loss function is still too high to efficiently fine-tune the pre-trained model. To solve the problem mentioned above, we use an efficient deep learning backbone as our person re-identification module assisted by the synthetic data to get the illumination invariant property. In order to perform high-speed training, a novel loss function, called adaptive weighted clustering loss, is used to replace the traditional metric learning loss function.

## Literature Review

We first give a brief review of human detection method mentioned in Section 1.2.1, followed by different methods for person re-identification in Section 1.2.2. At last, domain adaptation is introduced in Section 1.2.3.

### Human Detection

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) whole image | (b) background image | (c) foreground result |
| Figure 1‑2 The background subtraction result [7] | | |

Due to the requirement of various applications, human detection has raised lots of attention. According to the input data, human detection can be categorized into two aspects: (1) depth image-based human detection and (2) RGB image-based human detection.

For depth image, some methods [7-9] use background subtraction to show the foreground images, followed by use of a classifier to achieve human detection. Figure 1‑2 shows the background subtraction result, which requires background image initialization before background subtraction can be performed, and the foreground image (see Figure 1‑2 (c)) is generated by element-wise subtraction of the whole image (see Figure 1‑2 (a)) from the background image (see Figure 1‑2 (b)).

|  |
| --- |
|  |
| Figure 1‑3 The network architecture of RCNN [10] |

The above method is also suitable for handling RGB images. However, background subtraction cannot be applied to cluttered and mutative environment, which elevates difficulty when it comes to application. Dalal *et al.* [11] proposed HOG-based human detection method which can detect human even in cluttered backgrounds under varying illumination. Felzenszwalb *et al*. [4] trained multiple models to solve the problem where features are difficult to be extracted due to perspective change, and the performance has significant improvement over [11].

Girshick *et al.* [10] proposed RCNN and improves mean average precision (mAP) by more than 30% on VOC2012 [6]. Later, Girshick *et al.* improved their model [12] by proposing Fast-RCNN, which achieve higher accuracy and faster operating speed relative to RCNN. Because of the success of RCNN family, people began to focus on deep-learning based object detection which can also perform human detection.

However, the speed of RCNN family is too slow to be realistic for real-world applications. Redmon *et al*. [13] designed a novel architecture, You Only Look Once (YOLO) Network architecture, which can perform multiple object detection simultaneously. Due to its ability to achieve real-time detection, many researches use [13] as their model preprocessing to crop human or objects. However, there still remains a serious disadvantage for [13, 14], due to the multi-task properties the model has to predict object location and object classes simultaneously, which may degrade the model performance significantly.

The main difference between YOLO and YOLO-v2 is that, the latter uses anchor boxes to assist bounding box regression, which refers to the idea of Faster-RCNN [15]. If our model does not have any prior knowledge, it is hard to produce bounding boxes accurately. Due to the above reason, Faster-RCNN introduces anchor boxes which are predefined bounding boxes, and the work of our model is to fine-tune the size and the offset of these anchor boxes.

Note that YOLO and YOLO-v2 divide the image into grid cells. Each cell divided by YOLO predicts two bounding boxes and their respective classes, but each cell divided by YOLO-v2 predicts nine anchor boxes instead as well as their specific classes, as shown in Figure 1‑4.

|  |
| --- |
|  |
| Figure 1‑4 The difference between YOLO-v1 and YOLO-v2 |

From Figure 1‑5, YOLO-v1 uses fully connected layers to perform bounding box regression, which not only requires the input size of the model to be fixed but also uses more parameters to perform regression, where the latter is because the number of parameters of the fully connected layers is equal to the product of the number of prior unit and the number of next unit. YOLO-v2 uses the convolutional layers to replace the fully connected layer, which can effectively reduce the size of the model. Due to the characteristic of having fully convolutional, the Spatial Relationship of the features has been significantly improved and the accuracy of the model is also improved.

|  |
| --- |
|  |
| Figure ‑ The architecture of YOLO-v1 [13]. |

### Person Re-Identification

|  |
| --- |
|  |
| Figure 1‑6 The multi-channel part-based convolutional networks architecture [16] |

Person re-identification (Re-ID) problem depicts a scenario where given a query person image, one wishes to retrieve from the image gallery, presuming that all pedestrian image have already been detected by the human detection model. Since Gray *et al.* [17] proposed a challenging image dataset, Viewpoint Invariant Pedestrian Recognition (VIPeR), which contains more than six hundred pedestrian identities where photos are shot from two different cameras.

In pedestrian descriptions, color and shape of the pedestrian appearance is the most commonly used feature. Farenzena *et al*. [18] segmented pedestrian foreground from background, computed the weighted histogram, the recurrent high-structured patches, and the maximally stable color regions based on body part configuration, and achieved Re-ID by using the above features.

Li *et al*. [19] first introduced the CNNs method for Re-ID by using filter pairing neural networks. Cheng *et al.* [16] improve the triplet framework by proposing the multi-channel part-based convolutional neural networks for Re-ID as shown in Figure 1‑6. Some methods rely on the additional pre-trained model [18, 20-22] or additional dataset [23], Sarfraz *et al*. [20] proposed the method consisting of pose estimation and re-ranking, and showed that both the fine and coarse body pose cues are important for Re-ID. Cho *et al*. [21] introduced Pose-aware Multi-shot Matching (PaMM) that robustly estimates poses and efficiently conducts multi-shot matching based on the target pose information as shown in Figure 1‑7 . Su *et al.* [22] proposed a pose-driven deep convolutional model which is more suitable embedding the global human body and local body parts. Kalayh *et al.* [23] pre-trained their human semantic parsing model on Look into Person (LIP) [24] dataset, then combine the semantic human body part and high-level feature to perform combination. Even though [22] and [23] can achieve high performance on Re-ID, it requires dataset and 2D skeleton data which is not always available and it is easily influenced by noise while training.

|  |
| --- |
|  |
| Figure 1‑7 The network design for Pose-aware Multi-shot Matching [21]. |

Ristani *et al.* [25] proposed a simple but useful procedure, which uses Openpose [26] to detect human and implements feature extraction trained on Re-ID dataset [27]. They also proposed a novel loss function, named adaptive weighted triplet loss function, where general form of triplet loss can be described as follows:

|  |  |
| --- | --- |
|  | (1‑1) |

where is the margin, is the Euclidean distance, and denote positive and negative weights, respectively, and , and represent pools of positive sample and negative sample, respectively.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |
| Figure 1‑8 The different between traditional triplet loss and adaptive weights triplet loss [25]. | | |

Figure 1‑8 shows the difference between standard triplet loss and adaptive weighted triplet loss, and of the figure (a) depicts the feature distance between anchor and the negative samples, (b) illustrates the effect of standard triplet loss with weights, and , satisfying

|  |  |
| --- | --- |
| , | (1‑2) |

and (c) shows the adaptive weighted triplet loss, where weights are defined as follows:

|  |  |
| --- | --- |
|  | (1‑3) |
|  | (1‑4) |

When challenging samples appear in a batch., the method is able to achieve corresponding correction by penalizing samples that are challenging and lighten the weights on samples that are relatively easy to handle.

### Domain Adaptation

Deep learning methods have gained success because of big data. However, the domain shift issues significantly lower the performance for deep learning, due to its property of easily fitting on training data. Several works [28-30] apply the domain adaptation method not only to get the higher performance on training domain but also to improve the results on testing domain, where training domain and testing domain refer to the domains from different datasets.

|  |
| --- |
|  |
| Figure 1‑9 Domain adaptation using gradient reversal layer [31]. |

Ganin and Lempitsky [31] used the domain classifier and forced the classifier to predict where the data were from. To align two domains and learn the feature representation simultaneously, they further proposed gradient reversal layer to achieve the above mentioned goal by backpropagation as shown in Figure 1‑9. Since Goodfellow *et al.* [32] proposed the Generative Adversarial Network, adversarial training strategy has achieved significant success in other computer fields on transfer learning. Deng *et al.* [33] improved the cycle-GAN [34] and proposed the SPGAN which can preserve the identity information after the image-to-image translation.

## Contributions

In this thesis, we propose a learning-based person re-identification model which can extract the appearance feature from different lighting conditions. In addition, we try to use simple backbone which will be described in Section 2.2.2 and only use RGB image as input to identify human identification. The contributions of this work can be summarized in the following:

1. To verify the proposed method, we introduce a simple but efficient pipeline to collect our own Re-ID dataset, which contains more than 3,549 bounding boxes, 40 human identities taken from 5 cameras, inside the building of Depart of Computer Science and Information Engineering (CSIE) at National Taiwan University (NTU).
2. We propose a framework to learn an illumination-invariant feature from the synthetic dataset, which can reduce the collection effort. The illumination-invariant feature can encode human appearance feature from images taken under different lighting conditions into identical feature space, which refers to the illumination-invariant property.
3. We propose the novel clustering-based loss function for recognition task, which is called adaptive weighted clustering loss (AWCL). In contrast to the contrastive loss [35] and triplet loss [36], the proposed AWCL is more efficient than both loss functions, and the experimental results show that AWCL can help our proposed solver to achieve higher performance on two public datasets.
4. To assist the adaptive weighted clustering loss, we introduce a novel batch construction strategy, which is called hard clustering mining (HCM). HCM can make the model easier to discriminate hard negative identities.

## Thesis Organization

In this chapter, we have stated the introduction about human tracking and person re-identification, the challenges of this research, and the problem formulation. In addition, literature reviews go through the history of human detection and some related work of this research.

In Chapter 2, we build up some prerequisite knowledge related to our research. In this chapter, we will present some background knowledge we apply to our works. Due to the preprocessing for our dataset, the clustering method, namely, k-means process will be explained step-by-step. In addition, convolutional neural networks play an important role in this research, we will introduce some basic modules of CNNs, and the development history of the CNNs. Finally, we will illustrate the main idea of information retrieval (IR) and what is the function for IR in this thesis.

In Chapter 3, we first introduce the synthetic dataset which we that used to train our model. Then, we describe the way how to train our model with synthetic data and the framework to extract the appearance feature, which is invariant to illumination variations. After extracting the illumination-invariant features, we adopt the transfer learning technology, which is generative adversarial network (GAN) to diminish the domain shift between synthetic data and real data. Last, in order to overcome the disadvantage of the traditional metric learning, which suffers from high time complexity. In this thesis, we propose a novel clustering-based loss function, called adaptive weighted clustering loss, which not only can obtain equal accuracy like the metric learning but also be more efficient while training the model.

In Chapter 4, we describe how to collect our re-identity dataset in Dept. Of CSIE at NTU. Then, the process of data processing will be introduced step by step.

In Chapter 5, the experimental results consolidate the effectiveness of our proposed method. In order to ensure each proposed method is useful, we made a series of ablation studies which can verify the method can obtain the highly promising performance and overcome the disadvantage of the other methods.

In Chapter 6, we will conclude the contributions of this thesis. Finally, the future work will also be discussed.

# Preliminaries

In this chapter, some prerequisite knowledge is introduced. First we present cluster analysis and a classic clustering method, k-means. Secondly, we give background of deep neural networks which includes the convolutional neural network. Thirdly, we introduce the real-time pose estimation module which is used to be the preprocess of our collect dataset. Last, the information retrieval technology will be described clearly.

## Cluster Analysis and K-means

|  |  |
| --- | --- |
|  |  |
| 1. 2D data points without labels | 1. Clustering result |
| Figure 2‑1 Clustering 2D data points from 4 clusters | |

The goal of cluster analysis is a way to group distributed data according to their distance relationship with each other, the distance between the data in the same group (or cluster) have to be as close as possible. In the field of deep learning, the data points are often located in high-dimension space, and the correlation between each data points are defined by the distance assessment. Figure 2‑1 shows a clustering example.

There are several clustering methods such as hierarchical clustering, centroid-based clustering and density-based clustering in our prior knowledge.

|  |  |
| --- | --- |
|  |  |
| (a). Black dots are the data points and the color stars represent the random initial center point. | (b). After data point assignment result, each data points will be assigned to the nearest center point’s cluster |
| Figure 2‑2 The data point assignment in first iteration. | |

The HDBSCAN [37] is an improved version of DSCAN [38] by transforming into hierarchical clustering algorithm, the advantage of HDBSCAN is not required to pre-define the number of group which we want to clusters. Although it is the latest clustering algorithm, it is extremely sensitive to hyper-parameters, and the final number of clusters can be difficult to keep within a limit, resulting in complicated subsequent applications.

K-means [39] algorithm is the most popular centroid-based clustering algorithm due to its property of intuitiveness and its unsupervised approach such that human label is not required. k-means algorithm consists of the following steps.

Firstly, we have to pre-define the number of clusters K. Secondly, based-on value K we can random sample K points as the initialize center points, the dimension of each sample points are same as the data points. is denoted as follow:

|  |  |
| --- | --- |
|  | (2‑1) |

where is the dimension of each data points, ∂ and denotes the first initial time. The initialization position of the center point will significantly affect the operation time of the algorithm. Thirdly, calculate the distance of all data points to each center point and assign each data point to the nearest center point as shown in Figure 2‑2. After we get the new clustering result by using the random initial points, we can recalculate the new center points by using clustering result. The new center points can be written as:

|  |  |
| --- | --- |
|  | (2‑2) |

where means the data points which is assigned to cluster at , and means the total points in . Then repeat the third step until the center point no longer changes.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |
| (e) | (f) |
| Figure 2‑3 The k-means process. (a) means the given randomly initial center point, (b) represent the clustering results obtained by (a), we can re-calculate the new center points based-on clustering result after get (b) which is represented by (c) and so on. | |

Figure 2‑3 shows the general process of k-means.

## Convolutional Neutral Network

Since AlexNet [40] first used convolutional networks in image classification and achieved significant performance improvements in ImageNet [41]. Convolutional Neural Networks (CNNs) have achieved great amount of success in the field of computer vision, including image classification [40, 42], object detection [10, 12, 13, 15], action recognition, *etc*.

To simplify the explanation, we use image classification as an example. Traditional frameworks consist of hand crafted feature extractor like SIFT [3] feature or HOG [2] feature, and learnable classifiers such as Support Vector Machine (SVM) [5]. While the classifiers can be tuned by solving the objective function, we need to predefine the parameters of the feature extractors and it requires some domain knowledge and several trial and error for tuning the most suitable classifier’s hyper-parameters.

On the other hand, the CNN-based frameworks consist of learnable feature extractor and classifiers. Both the feature extractor and classifier can be learned through backpropagation in an end-to-end fashion, which makes it easier to learn how to extract the suitable and reliable features. Figure 2‑4 shows a standard CNN architecture for classification task.

|  |
| --- |
|  |
| Figure 2‑4 A standard classification task CNN architecture comprised of several convolutional layers and fully connected layers, the end-to-end training strategy can guarantee that CNN and fully connected layer are optimized simultaneously. |

The difference between traditional neural networks and convolutional neural networks is that the neural network uses node-to-node operations, while convolution neural network consists of the concept of spatial information. In order to achieve the above properties, the convolution network performs convolution of 2D images through the convolution kernel. Each kernel receives some inputs, performs a dot product and non-linearity activation is optional to increase the nonlinearity of the model. Zeiler *et al.* [42] introduced the visualize method which uses Deconvolution to visualize the feature map response, they also show that earlier layers extract low-level features, such as edge, corner and color, while following layers are responsible for extracting high-level features.

### Convolutional Layers

Convolutional layer is an important role in Convolutional Neural Network. Each convolutional layer has multiple convolution kernels which perform dot product with a specific input patch, the intensity of the feature can be obtained. Each layer can have different number of kernels to output differ in channel (or depth) dimension.

|  |
| --- |
| ../../螢幕快照%202019-03-22%20下午11.27.32.png../../螢幕快照%202019-03-22%20下午11.28.14.png../../螢幕快照%202019-03-22%20下午11.28.26.png |
| Figure 2‑5 The standard convolution operation with stride equal to 1 |

During the forward pass, a two dimensional convolution kernel will perform dot product with the input feature map as shown in Figure 2‑5. When training from scratch, the weight of each convolution kernel can be randomly initialized, and the weights are updated through backpropagation which is calculated by the gradients of the loss function.

Since 2014 Simonyan *et al.* [43] proposed VGG Network, and the difference between AlexNet [40] and VGG is that VGG uses the deeper architecture given that the deeper model can produce higher accuracy. Figure 2‑6 shows the difference between AlexNet and VGG16 architecture.

|  |
| --- |
|  |
| 1. AlexNet architecture [40] |
|  |
| 1. VGG16 architecture [43] |
| Figure 2‑6 The network design for AlexNet and Standard VGG16. |

### Residual Network

According to Universal Approximation Theorem [44], we know that we can use a feedforward neural network to describe any equation. The most acceptable way is allowing the model continue to deepen.

To expand the capacity of VGG modules, more and more researches try to use more convolution layers to test the limit of deep learning. Vanishing gradient problem [45] makes this trend obstructed, this may completely prevent the neural network from further training. Since the backpropagation computes gradients by the chain rules, the gradient of the front layers will be decreased more than the following layers.

|  |  |
| --- | --- |
|  |  |
| 1. Traditional Convolution Block | 1. Residual Block |
| Figure 2‑7 The difference between traditional convolution block and residual block. | |

To solve the problem mentioned above, Kiming *et al.* [46] proposed the Residual Network (ResNet) which is feedforward neural networks by adopting shortcut connections which allows gradients to not vanish during conduction due to too many differentiations, by doing so the deeper network is again trainable. Figure 2‑7 illustrates the shortcut connection performing identity mapping and shows the difference between traditional convolution block and residual block. The residual block can be written as follows:

|  |  |
| --- | --- |
|  | (2‑3) |

The ResNet-50, as shown in Figure 2‑8, has 50 convolutional layers formed into 5 modules, which are equipped with short connections. The ResNet-50 is often used as the backbone of deep learning models, its powerful ability of extracting high-level features are shown in many common datasets. In this thesis, we also use ResNet-50 without fully connected layer to extract the high-level features further details would be revealed in Section 3.1.2.

|  |
| --- |
|  |
| Figure 2‑8 The family of ResNet [46], each layer contains multi residual block. The model is named according to the number of layers it has. For example, we named 50-layer as ResNet-50.   |  | | --- | |  | | Figure 2‑9 The architecture of the Openpose [26] | |

## Real-time Pose Estimation Module

In recent years, many human detection methods have been developed, and Openpose [26] is a most famous among them all. Openpose is the fully convolutional architecture, which can tolerate different size of input image. Due to the above property, it becomes the most robust deep learning-based human detection framework. Figure 2‑9 shows the architecture of the Openpose. Besides, Openpose uses the bottom-up human pose estimation method, and it also includes part affinity fields to associate body parts with individuals in the image. The difference between bottom-up and top-down pose estimation is that, top-down methods need to use human detection as the preprocessor and perform pose estimation based on the detection patch. However, bottom-up have the potential to decouple runtime complexity from the number of people in the image without any preprocessing.

|  |  |
| --- | --- |
|  |  |
| (a) The Input image | (b) Body part confidence maps |
|  |  |
| (c) Part affinity fields | (d) Pose estimation result |
| Figure 2‑10 Pose estimation process for Openpose[26] | |

Figure 2‑10 show the process for Openpose [26], the model can generate predictions of body parts (see Figure 2‑10 (b)) and part affinity fields (see Figure 2‑10 (c)) simultaneously, and then associate body parts with maximum part affinity fields between two joints by using greedy algorithm (see Figure 2‑10 (d)).

The main advantage of Openpose is that it can perform real-time application. Due to the bottom-up property, Openpose does not need human detection as a preprocessor; therefore, it is more suitable application wise [25] than top-down methods.

|  |
| --- |
|  |
| Figure 2‑11 The idea of the information retrieval on Person Re-identification |

## Information Retrieval

In this section, we give a brief introduction about Information Retrieval (IR), which contains a large variety of applications, such as search engines, media search and digital libraries. The goal of IR is to search for information similar to a given query target. In this thesis, information retrieval is applied in the part of person re-identification. The task of person re-identification is to retrieve images from the gallery image when receiving a query image as shown in Figure 2‑11.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2‑1 Confusion matrix | | | |
|  | | Prediction | |
| True | False |
| Ground truth | True | True Positive | False Negative |
| False | False Positive | True Negative |
|  |  |  |  |

Mean Average Precision (mAP) is a very popular performance metric in information retrieval. Before explaining mAP, knowledge about precision and recall is required. Both of them are used to evaluate the performance of the IR algorithm. The definition of the precision and recall are listed as follows:

|  |  |
| --- | --- |
|  | (2‑4) |
|  | (2‑5) |

where , and denotes true positive, false positive and false negative which are obtained from the confusion matrix. Table 2‑1 gives an example of the confusion matrix.

The mean average precision calculates the area under precision-recall curve, and it is a popular way to evaluate the overall performance of IR algorithm. More definition about mean average precision will be described in 5.3.3.

# Person Re-Identification

In Section 3.1, we first describe how to learn the Illumination-Invariant Feature (IIF) that is transferred from appearance feature in different lighting conditions to a shared high-level feature space which is an illumination-invariant domain. In order to overcome the disadvantage of the contrastive loss [35] and triplet loss [36], we have propose the clustering-based loss function which will be described in Section 3.2.

## Learn an Illumination-Invariant Feature

Training deep neural networks requires on collect large amount of labelled data, which is high expensive and time consuming. In our case, learning an illumination-invariant feature has to collect the data from different illuminations and environments. It is very hard and expensive to manually collect and label the data. In order to avoid spending too much time on collecting the corresponding training data, several works [47-49] train on synthetic data, and bring rival results while evaluating on real data. Inspired by the above works, synthetic data become the solution in our thesis that need to collect a large amount of training dataset.

In Section 3.1.1, we first describe the synthetic dataset we used to acquire the illumination-invariant property. Then, the process of training model using synthetic data is shown in Section 3.1.2. However, the extracted illumination-invariant feature (IIF) is distilled from the synthetic dataset, which will cause the properties of IIF cannot be mapped to the real data directly. Therefore, we introduce the domain adaptation technology to transfer the knowledge from synthetic domain to real domain in Section 3.1.3.

### Synthetic Dataset

According to the property of the dataset, we have choose SyRI dataset [49], which provides 100 human identities under 140 illumination conditions, and randomly image every human from 4 different viewpoints under each illumination resulting in 56000 images (), to be used for training our model. The dataset uses multiple High Dynamic Range Image (HDRI) environment maps which accurately simulate real indoor and outdoor lighting. Moreover, the 3D virtual humans are designed with *Adobe Fuse CC*, which can provide 3D content including body scan of real people with customizable body, and rendering by using the *Unreal Engine 4*, which is a suite of integrated tools for game developers to design and build games, simulations, and visualizations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | View1 | View2 | View3 | View4 |
| Id1 |  |  |  |  |
| Id2 |  |  |  |  |
| Figure 3‑1 The image capture from SyRI dataset. | | | | |

Figure 3‑1 shows some sample images of the SyRI dataset. As we can see, the first row (upper) shows the same identity but subject to different camera position setting.

When training the model, we have to resize all of the image to the same size which enables the GPU to assist in parallel operations in one batch. By doing so, this is the reason why the first image in first row seems fatter than the other image, but actually all of them in the first row share the same identity label.

### Learn from synthetic data

Inspired by [50-52], we proposed the network which consists of two parts, encoder and an identity classifier , as shown in Figure 3‑2. We intend to let the same person from different illuminations to share the same high-level feature space. To achieve this goal, we have to build a powerful feature extractor, which can extract the robust features. As described in Section 2.2, Convolutional Neural Networks (CNNs) can play this role by training our model with big data.

|  |
| --- |
|  |
| Figure 3‑2 Architecture for learning appearance feature. |

Due to the above reason, the encoder has been designed as a CNN-based structure, which extracts the appearance feature from image . Then, the identity classifier discriminates the feature extracted from encoder , which may enlarge the distance of feature vectors between different persons and diminish the distance of feature vectors belonging to the same person under difference illumination conditions in high dimension space.

For encoder , we utilize the residual network [46] which contains 50 layers, called “ResNet-50” as described in Section 2.2.2. In order to make the model work on our problem, the last fully connected layer is removed and is replaced by “Batch Normalization Layer” [53], resulting in the appearance feature which is a vector of 2048-dimension vector generate by ResNet-50. The identity classifier is designed as a single layer fully connected network with activation layer, softmax function, which represent the probability of each class.

To learn the discriminative feature, we incorporate the other classification-based methods [50-52], using standard cross-entropy as the loss function as shown below.

|  |  |
| --- | --- |
|  | (3‑1) |
|  | (3‑2) |
|  | (3‑3) |

|  |
| --- |
|  |
| Figure 3‑3 Visualization of illumination-invariant feature extracted from SyRI dataset. Each cluster represents the features with same identity from different lighting. Due to the limitation of the palette, same color may appear multiple times. |

where is the number of identity labels, is the -th element of , which indicates the softmax activations of the identity classifier. is the identity label for the image , and denotes Iverson bracket. Note that the image is from SyRI dataset.

|  |
| --- |
|  |
| Figure 3‑4 Domain shift visualization which is performing dimensionality reduction by using t-SNE. Orange points is illumination invariant feature extracted from real data, blue points are illumination invariant feature extracted from synthetic data. |

After training, we use t-distributed stochastic neighbor embedding (t-SNE) [54] to project the high-level illumination-invariant feature extracted from SyRI dataset to a 2-Dimensional space. Figure 3‑3 shows that each cluster represents the feature of the same identity but under different illuminations. Figure 3‑3 also shows that Encoder is able to extract similar features with same identity from different lighting, which meets our objective.

### Domain Adaptation by Adversarial Learning

After we training the illumination invariant encoder by using SyRI dataset, we found that our model suffers from the domain shift issues which let invariant encoder cannot accommodate real data. Figure 3‑4 shows the influence of the visual data bias problem, which refers to that is a gap exists between synthetic data (blue points) and real data (orange points).

Training the model by using large real dataset which has the same property as synthetic data can efficiently solve the domain shift problems. However, it’s hard and expensive to collect the new dataset, and collecting a new dataset does not meet the original intention of using synthetic dataset. Therefore, domain adaptation is the suitable solution in our case, which makes the distribution of synthetic data more similar to that of real data. Inspired by [31, 55], we use the popular domain adaptation technology, namely, adversarial discriminative domain adaptation, which uses the standard generative adversarial network (GAN) to achieve domain adaptation property.

|  |
| --- |
|  |
| Figure 3‑5 The complete architecture for learning illumination-invariant feature with domain adaptation. The blue line means the flow for synthetic data and the red line is for real data, respectively. |

In order to make sure that our model processes the illumination-invariant property, firstly we train our encoder network by using synthetic data as described in Section 3.1.2 and secondly we consider to be domain-invariant if there exists a powerful domain classifier but it cannot clearly discriminate the features extracted from real data and those from synthetic data. From the above observation, we utilize a fully connected network as our domain classifier, and each domain (Target and Synthetic) has its own Identity Classifier as shown in Figure 3‑5.

As we can see, the inputs of our complete architecture not only accept synthetic person images but also real data . It is worth mentioning that the real data are not collected under dramatic variations in illumination, and then the proposed model has to learn the illumination-invariant property only through the synthetic data collected under intense changes in the lighting.

To distinguish the domain of the features, the objective function for the domain classifier is expressed as follows:

|  |  |
| --- | --- |
|  | (3‑4) |

In order to minimize domain loss, , when , the Domain Classifier has to predict 0, and to predict 1 when . When optimizing Eqn. (3‑4), we freeze the Encoder to ensure Domain Classifier can effectively learn how to distinguish the feature extracted from real data and that from synthetic data. In addition, to freeze the Encoder is an important step. If we perform the backpropagation and do not froze the Encoder , the gradient will cause Encoder to extract the domain variant feature.

Besides, we have to design an additional loss function to let the Encoder be able to compete with the Domain Classifier , the loss function can be written as follows:

|  |  |
| --- | --- |
|  | (3‑5) |

In Eqn. (3‑5), we freeze the Domain Classifier , only update the parameters of Encoder , which forces the Encoder to know how to confuse the Domain Classifier and complete the task by itself.

## Assist by Clustering

As described in [56], the deep learning based features need to be not only separable but also discriminative in the recognition task, and such idea is also suitable for Re-ID. Several works [25, 28, 33, 57] used the triplet loss or contrastive loss to enlarge the distance among feature vectors corresponding to different identities and diminish the distance of feature vectors corresponding to the same identities, which is useful but time consuming while training the model. The general of the triplet loss as described in Eqn (1‑1) and the contrastive loss can be written as follows:

|  |  |
| --- | --- |
|  | (3‑6) |

where represents the batch size or number of samples, and and are margins, and . The disadvantage of the contrastive loss is that the time complexity of the contrastive loss is equal to and triplet loss is , and since the margins are fixed, the embedding feature space of the contrastive loss cannot produce distortion, which indicates that hard identity mining [25] needs to be incorporated to let the contrastive loss avoid overfitting. Moreover, the disadvantage of the triplet loss is that the margin is only used to force the distance between sample and the positive as close as possible whereas the distance between sample and negative as far as possible.

Triplet loss still can be used to assist solving re-ID problem because there must exist a sample which can pair the query while evaluating the model. However, in real-world application, we are not sure that the query already exists in our database, which means that we have to define the margin to discriminate positive sample and negative sample.

### Clustering Loss

According to what is mentioned above, we propose the clustering-based loss function, which not only retains the properties of the distance-based methods (triplet loss and contrastive loss) but also is faster than the distance-based methods. Bert *et al.* [58] adopted the discriminative loss function to enforce the network to clearly discriminate different objects in the same image and make the pixels from the same object closer. This idea is the same as triplet loss but computation cost is lower. However, the regularization term of the discriminative loss might lead to confusion in the training process because the cluster center of the identities should not be close to zero.

Inspired by [58] and k-means [39], we modify the discriminative loss and propose the clustering-based loss function, called clustering loss (CL). The total loss function for discriminating appearance feature is defined as:

|  |  |
| --- | --- |
|  | (3‑7) |

|  |
| --- |
|  |
| Figure 3‑6 The flow of the proposed method. |

However, if classification loss and clustering loss are backpropagated to the same feature vector, the gradient of two loss function may confuse model [53, 59] and make performance sharply drop. In order to solve this problem, we apply the batch normalization between backbone and classifier, and the flow of the loss function is shown in Figure 3‑6.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
| Figure 3‑7 The idea of clustering loss’s variance term (a) and distance term (b) | |

The clustering loss (CL) is designed as follows:

|  |  |
| --- | --- |
|  | (3‑8) |
|  | (3‑9) |
|  | (3‑10) |
|  | (3‑11) |

where and are hype-parameters, and are the margins, represents the feature before batch normalization extracted by encoder by input image which belongs to class p, and and denote the batch construction parameters.

In order to retain the properties of the distance-based methods, the clustering loss contains two terms, where the first one is the variance term in Eqn. (3‑9) which is used to diminish the distance between features and center point from the same class in one batch smaller than (see Figure 3‑7 (a)), and the idea of the variance term is to try to diminish the distance between anchor and positive sample; the second term is the distance term in Eqn. (3‑10) which is used to enlarge the distance between each center greater than (see Figure 3‑7 (b)), and the idea of the distance term is to try to enlarge the distance between anchor and negative sample.

For batch constructions, we adopt the batch, which is introduced by Hermans *et al.* [60]. In each batch, there are images for each of identities. batch can increase the number of positive and negative samples in a batch. The time complexity of the proposed CL is , the time complexity of the triplet loss is , and the time complexity of the contrastive loss is .

|  |  |
| --- | --- |
|  |  |
| (a) variance term | (b) adaptive weighted variance term |
| Figure 3‑8 The different between uniform weighted and softmax weighted, the length of the arrow represents force. | |

### Adaptive Weighted Clustering Loss

The proposed Clustering Loss can reduce lots of computation time due to the low time complexity. However, the variance term is uniformly weighted for each feature point, which may cause the model not only to abandon difficult samples but also to make training efficiency drop dramatically.

To improve the performance of the proposed clustering loss, we add the weights on the variance term in Eqn. (3‑9). Inspired by [25], we define the weights to give little importance to the easy samples and emphasize the hard samples, and name the modified variance term for adaptive weighted variance term. The improved loss function, Adaptive Weighted Clustering Loss (AWCL), can be described as follows:

|  |  |
| --- | --- |
|  | (3‑12) |
|  | (3‑13) |
|  | (3‑14) |

We assign the weights using the softmax as shown in Eqn. (3‑14), which can make the model emphasize the most difficult ones. Note that, the weighting term is active only when the distance between cluster center and feature is greater than . After applying the weighting, the adaptive weighted variance term’s physical meaning is shown in Figure 3‑8.

### Hard Clustering Mining

The other improvement of this thesis is in the procedure of selecting difficult clusters. Ristani *et al.* [25] said that without the selection procedure, it is hard to see difficult samples for each identity. To increase the chances of seeing hard negatives, [25] propose the image-based difficult identities selection, which is used to make sure that the distance between difficult samples can be enlarged by using the triplet loss, called Hard Negative Mining (HNM). However, HNM is not suitable for our proposed methods because clustering loss is used to enlarge the distance between cluster centers but not the distance between two different identities. According to the above idea, we propose the Hard Clustering Mining (HCM) to ensure that if the centers of two clusters are too close to each other, then that enlarged distance of the two centers is by using the clustering loss. The HCM samples hard negatives more frequently and preserves the half of random negative sample to ensure the randomness of the batch.

The hard samples can be constructed after training the network for few epochs. Figure 3‑9 illustrates the idea of the hard clustering mining, and the number represents different clusters, *e.g.*, number 1 represents the anchor. During origin batch construction, the negative samples will be randomly sampled, which causes the distance term of the adaptive weighted clustering loss hard to enlarge the distance between the anchor and hard negative. By contrast, the proposed hard clustering mining ensures that hard negative sample can be chosen.

|  |  |
| --- | --- |
|  |  |
| (a) Origin batch construction perform random sample. | (b) Hard clustering mining can choose the hard sample. |
| Figure 3‑9 The difference between origin batch construction and HCM | |

# ACL Re-Identification Dataset

In this thesis, we collect the dataset in order to fine-tune the model from the dataset- domain to our environment-domain. This dataset contains 40 identities who are the members in our laboratory. Note that when the member changes his/her clothing, it represents a new identity. As a result, our method is more concerned about the wearing rather than the face identity.

## Environment setting

|  |
| --- |
|  |
| Figure 4‑1 The camera topology in CSIE-Der Tian Hell’s second floor |

To collect the dataset, we choose five different kinds of environment in Dept. of CSIE at NTU and set the cameras at various places such that their FOV(field of view) are almost not overlapping. Figure 4‑1 shows the camera topology in our setting. As we can see, three cameras are shooting toward the corridor (2, 3, 4), camera1 is used to capture the indoor environment, and the camera5 is shooting towards the stairs. We give an example image of each camera for our dataset as shown in Figure 4‑2. Each video stream is recorded in the normal lighting while the subject walks around arbitrarily to capture different angles. In this way, we can make sure that our deep learning person re-identification model can handle different human orientations as shown in Figure 4‑3.

|  |  |
| --- | --- |
|  |  |
| (a) camera1 | (b) camera2 |
|  |  |
| (c) camera3 | (d) camera4 |
| (e) camera5 | |
| Figure 4‑2 The example of each cameras | |

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |
| (e) | (f) |
| Figure 4‑3 The Example of different pose and different human orientation | |

## Preprocessing

After collecting the dataset, we preprocess our dataset with the following three steps as shown in Figure 4‑4. Firstly, due to our own dataset, called ACL-reID (ACL is acronym for Advanced Control Lab) is collected from different cameras which may cause the video streams to fail to have the same image size. Thus, we resize the image to the size of , and use Openpose, which is described in Section 2.3 to perform the pose estimation. Since all of the videos in the dataset we record only contain one human, if there is more than one human or none in the result of human pose estimation of an image, we ignore the results of this image, which means we abort this image. Secondly, Openpose only returns the set of the joint positions, and the values are the ratio rather than real position. As a result, we have to multiply the ratios and the image size to acquire the real positions, and afterwards we can get the bounding box location with the help from the following padding transformation equation:

|  |  |
| --- | --- |
|  | (4‑1) |
|  | (4‑2) |
|  | (4‑3) |
|  | (4‑4) |

where and are the padding values which can prevent the bounding box from being too tight to crop the entire human, denotes the set of X-axis coordinates of each joint, denotes the set of Y-axis coordinates of each joint, and and represent the width and height of the image. Figure 4‑5 shows the influence of the padding. In the figure we can find out that, if we do not apply the bounding box padding, we may lose some human information such as shoes and hairstyle. After getting the bounding box location, we crop the bounding box and resize all of the region of interest to to form our new re-ID dataset.

|  |
| --- |
|  |
| Figure 4‑4 Pipeline of dataset preprocessing |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |
| Figure 4‑5 The difference between without padding (b) and with padding (c). | | |

Last, we define the “junk” image if the union area of Openpose detection (see Figure 4‑6 blue box) and the cropped human position (see Figure 4‑6 red box) is less than 40%, which means the cropped human is too small to become the retrieval image. Due to the “junk” images that make the model confuse and drop the performance, our dataset abandons the “junk” images to ensure the high quality.

When naming the image in the dataset, we follow the rule of Market-1501 [61] and DukeMTMC-reID [27]. Each image is named as “\_\_”. In order to reduce labeling time, we assume that there is only one man in the video when collecting the videos and the person does not change the apparel during the video. Therefore, we name the video as “\_\_”. This strategy allows us to name our region of interest easily by using the name of the video and image number in the video.

|  |
| --- |
|  |
| Figure 4‑6 The junk image sample, blue box represents the Openpose detection result, the red box denotes the cropped region which is padding from the blue box. |
|  |

To quantify the fine-tuning result on our environments, we design an experiment for cross-camera person re-identification. We split the dataset into two parts, the training part contains 2877 images from 30 identities, and testing part contains 622 images from the other 10 identities. For the purpose of the query set composition, we sample 1 image from different identities and different cameras from the testing set. The data distribution is shown in Figure 4‑7.

|  |
| --- |
|  |
| 1. Training distrubition |
|  |
| 1. Testing distrubition |
| Figure 4‑7 The distribution of the ACL-reID data |

# Experiments

In this chapter, we introduce the environment setting in this thesis. The following is the implementation details in this thesis. Before the experimental results, we give the introductions of two public person re-identification datasets and the two evaluation metrics. In order to proof that every part of proposed method is useful, we design a series of the ablation studies. In the end, we compare our method with various state-of-the-art ones on two public datasets, including Market-1501 [61] and DukeMTMC-reID [62]. The quantitative experimental results confirm the advantages of the proposed adaptive weighted clustering loss and illumination-invariant feature.

## Configuration

The specification of our experiment environment is listed below in Table 5‑1. In this thesis, we utilize Pytorch as our deep learning develop API. Our model is trained on the personal computer equipped with NVIDIA GeForce GTX 1080ti GPU, which has 11G memory and afford the training of our model.

|  |  |
| --- | --- |
| Table 5‑1 Specification of our experiment environment | |
| Central Processing Unit | Intel Core i5-7400 |
| Memory | 16G |
| Operating System | Ubuntu 16.04 LTS 64-bit |
| Deep Learning API | Pytorch1.0.1 |
| Graphic Processor Unit | NVIDIA GeForce GTX 1080ti |
| Cuda version | 9.0 |
| Cudnn version | 7.0 |
| Programming Language | Python 3.5.2 |

## Implementation Details

The implementation details including network designs, parameters setting, and some training tricks are mentioned in this section.

### Network design

In this part, we describe more details about our propose architecture. The feature extraction (Encoder ), synthetic identity classifier , domain classifier , and the target data identity classifier are shown in Figure 3‑5.

|  |
| --- |
|  |
| Figure 5‑1 The hold architecture of the Encoder |

For the Encoder , we utilize the “ResNet50” [46] which contains 50 layers and the last fully connected layer is replaced with a “Global Average Pooling” following by “Batch Normalization Layer” [53] as shown in Figure 5‑1.

Not only do the batch normalization normalize the feature to the hypersphere, but also ensures the adaptive weighted clustering loss will not counteract the effect of classification loss.

As for the identity classifier and , we design a simple one-layer fully-connected neural network with dropout mechanism. We feed it with 2048-dimension latent feature , and there are output units, where the output size would be 751, 702, 100 if dataset are Market-1501 [61], DukeMTMC-reID [27, 62] , SyRI dataset [49], respectively.

Last, for the domain classifier , a 2-layer fully-connected neural network is designed. The first layer comprised with 512 units followed by a ReLU [63] and the last layer aims to do classification using one unit with sigmoid function.

### Training Details

The ResNet50 is initialized with the weights which are pre-trained on Image-Net [41] and the input image first goes through data normalization using and . For the batch construction, we follow the idea of [60] while performing the batch construction. In each batch, there are image samples for each of identities. Many results have showed that batch has an excellent performance in similarity-based ranking because it can easily generate the positive and negative samples. In addition, we augment the training image with random crops, horizontal flips and perspective transformation to gain additional viewpoint invariance. For robustness to various resolutions, we apply the Gaussian blur of varying .

As shown in Figure 5‑2, we design the four-stage training process. The first stage is used to learn IIF encoder, the second and third stage are used to enable the encoder to make real data and also to have IIF property, and the last one is for learning discriminative feature which can distinguish different identities.

For the training, we set the hyper-parameters of the adaptive weighted clustering loss margin, , , , , the parameters of the batch construction , and the input resolution is resized to . Besides, the model is optimized by the stochastic gradient descent (SGD) algorithm [64] with the learning rate is initialized with , and the learning rate decays in half every 10000 iterations. In the experiments of hard clustering mining, we first train our model by 5000 iterations. Then perform hard clustering mining to find the top 15 closest negative identities for each identity, hereby called hard negative cluster set. Moreover, we reset the hard negative cluster set every 2000 iterations to ensure each cluster can stay away from others.

|  |
| --- |
|  |
| Figure 5‑2 The training stages on our proposed method |

## Person Re-Identification Dataset

To evaluate our propose method, we evaluate it on two public datasets, Market1501 [61] and DukeMTMC-reID [27, 62], which are the most famous person re-identification datasets. Both datasets contain large quantity of images, each with camera label. After introducing the dataset, we define the two evaluation metrics for the person re-identification task.

### Market-1501 Dataset

|  |
| --- |
|  |
| Figure 5‑3 The different identities collected from different cameras. |

This dataset [61] contains more than 30,000 annotated bounding boxes, and there are 1501 identities collected from 6 different cameras on Tsinghua University, China. The images in datasets are produced using the Deformable Part Model (DPM) [4] as pedestrian detector. The dataset is split into two parts, training part contains 12,936 images from 751 identities and testing part contains 19,732 images from 750 identities. Figure 5‑3 demonstrates some images in Market-1501 dataset. As we can see, the second row shows different persons but wearing same shorts, which make this dataset quite challenging since the model need to discriminate different identities who wear the same color’s short but with different accessories. The biggest difference between this dataset and the prior person re-identification datasets is that prior datasets have only one ground truth and one query image for each identity (close environment) on testing set, but Market-1501 is collected in an open system, where each identity may have multiple images under each camera.

Besides, this dataset splits the bounding boxes as two categories. The “good” image is when the ratio of the overlapping area to the union area between predicted boxes and hand-drawn boxes is greater than 50%, and them the DPM bounding box is marked as “good”. On the other hands, if the ratio is less than 50%, the bounding boxes is marked as the “junk”. Since the “junk” bounding boxes may distract the model, during evaluation we can ignore the influence of the junk bounding boxes. Figure 5‑4 shows the sample of junk image. As we can see, some images only contain the part of the human, or there is no the human in the images.

|  |
| --- |
|  |
| Figure 5‑4 Sample images of the distractor dataset (junk) [61] |
|  |

### DukeMTMC-reID Dataset

DukeMTMC-reID [62], a subset of the DukeMTMC dataset [27] for image-based re-identification, which is a large-scale tracking and person re-identification dataset recorded at the Duke University. It contains almost 1400 person identities across 8 static and synchronized cameras, which are mounted outdoors on the Duke University campus without overlapping. Figure 5‑5 illustrates the camera setting for DukeMTMC-reID, and there exist 8 cameras placed at different positions with almost no overlapping. In addition, DukeMTMC contains many kinds of different subsets, such as DukeMTMC-attribute and DukeMTMC-Pose. Moreover, for the evaluation metric of the dataset, DukeMTMC-reID, is the same as that for Market-1501, which uses Cumulative Match Characteristic (CMC) curve and mean average precision (mAP).

|  |
| --- |
|  |
| Figure 5‑5 The camera position setting for DukeMTMC-reID [27] |

|  |
| --- |
|  |
| Figure 5‑6 The image distribution on DukeMTMC-reID training set. |

This dataset provides three folders. Firstly, the training set contains 702 identities and totally have 16,522 images which are the humans cropped from the dataset, DukeMTMC. Secondly, the query set has 2,228 images of the other 702 identities. Thirdly, the testing set, contains 702 identities + 408 distractor and 17,661 gallery images, where distractor represents the junk image. Figure 5‑6 represents the data distribution on DukeMTMC-reID training set. As we can see, the median of images per ID is 20. But some ID may contain lots of images, like ID 5388 contains 426 images.

|  |
| --- |
|  |
| Figure 5‑7 The Cumulative Match Characteristic (CMC) curve. |

### Evaluation Metrics

Person Re-ID cannot use accuracy as the evaluation metric due to the identities in testing set haven’t been seen in training set. Hence, the person Re-ID is a sorting problem that need to calculate similarity of all photos to the query image.

In this thesis, we evaluate our method by using mean average precision and Cumulative Match Characteristic (CMC) curve. The CMC curve comprehensively reflects the performance of the information retrieval module. The larger the area under the curve (AUC), the better the performance of the classifier as shown in Figure 5‑7. The evaluating indicator of CMC curve is the same as the common deep learning indicator, which is top1 error and top5 error, but there remains slight difference from CMC curve using the rank1 and rank5, listed in the following equation:

|  |  |
| --- | --- |
|  | (5‑1) |
|  | (5‑2) |

In addition, mean average precision is very sensitive to the ranking of retrieval results. The relevant object that are ranked higher contribute more to the average than the relevant object that are ranked lower. The definition of the mean average precision can be written as follows:

|  |  |
| --- | --- |
|  | (5‑3) |

where is the number of the query images, represents the number of the gallery images with the same identity to the query , denotes the number of all the gallery images, illustrates the precision at rank for the *-th* query, and means the identity of *-th* prediction.

## Cross-Illumination Classification Result

Due to [49] do not define how to evaluate the performance of the cross-illumination classification result. To quantify the result of the illumination-invariant feature, we split the SyRI dataset into training set and validation set. And we assume that the model has to classify the person identification from the unseen illumination and environments.

|  |  |
| --- | --- |
| Table 5‑2 Identity classification accuracy on SyRI dataset. | |
| Training set | Testing set |
| 98.4% | 96.8% |

All 140 illuminations are split into 112 training illumination and 28 testing illumination. Thus, we have 44,800 training image and 11,200 testing images. Note that the training data and testing data share the same identity set but under different illuminations, which is used to verify that our model can recognize the person from unseen illumination. During testing, we use the Encoder and Identity Classifier which is trained on training set as described in Section 3.1.2.

The result is listed in Table 5‑2. Although the testing set is captured from different illuminations, we can still achieve high accuracy by 96.8%, which represents the robustness of the Encoder to varying illuminations.

## Person Re-Identification Result

To validate the performance of our proposed method for person re-identification, we first do the ablation studies to ensure each part of our method is effective. Secondly, we compare the performance with several state-of-the-art methods which are published on the top computer vision conferences.

### Ablation study

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 5‑3 Comparison of different kinds of loss and the proposed loss function on the Market-1501 [61] and DukeMTMC-reID [62]. \* denotes the data of the experiment results are copied from the paper. | | | | | |
|  | Market-1501 | | DukeMTMC-reID | |  |
| Methods | Rank-1 | mAP | Rank-1 | mAP | Run time |
| baseline | 81.71% | 60.34% | 71.85% | 52.50% | 0.339sec |
| baseline  +  contrastive loss | 85.57% | 67.00% | 73.23% | 55.40% | 0.778sec |
| AWTL [25] \* | 86.94% | 71.76% | 75.31% | 57.28% | > 1 sec |
| baseline  +  AWCL | 88.33% | 72.71% | 74.23% | 57.90% | 0.359sec |

1. *Additional Loss Function*

For the novel adaptive weighted clustering loss function, we do the experiment on Market-1501 dataset and DukeMTMC-reID dataset to verify that the proposed loss function is more suitable than the contrastive loss function [35]. Table 5‑3 demonstrates the comparison of contrastive loss and the proposed loss function. The baseline model represents ResNet50 + fully connected layer with cross-entropy. The baseline + contrastive loss means using not only using cross-entropy but also contrastive loss to train the model. However, because triplet loss is too time-consuming to implement. As a result, we compare with the advanced triplet loss, called Adaptive Weighted Triplet Loss (AWTL) [25]. The run time (last column) represents how much time we take to train the model per batch. In addition, the table shows that, when we adopt adaptive weighted clustering loss and make a comparison with the baseline model (first row), we found out that it only takes additional 0.02 seconds to calculate the adaptive weighted clustering loss, but contrastive loss takes 0.439 seconds, which means the proposed adaptive weighted clustering loss is faster than contrastive loss 20 times. We observe that adaptive weighted clustering loss not only achieves higher accuracy than contrastive loss function but is more efficient than the contrastive loss and triplet loss.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 5‑4 Ablation studies of IIF. DA is acronym for domain adaptation. | | | | |
|  | Market-1501 | | DukeMTMC-reID | |
| Method | Rank-1 | mAP | Rank-1 | mAP |
| baseline | 81.71% | 60.34% | 71.85% | 52.50% |
| IIF w/o DA | 84.32% | 66.10% | 73.56% | 54.25% |
| IIF w/ DA | 86.01% | 68.35% | 75.62% | 55.65% |
|  |  |  |  |  |

1. *The Effect of the illumination-invariant property and domain adaptation*

In this subsection, we aim to analyze the effectiveness of our method through comparison with the baseline model as shown in Table 5‑4. The baseline is trained on both datasets without any additional efforts. IIF w/o DA represents the model which trains on the Synthetic data and gets the illumination-invariant feature (IIF) but without adopting domain adaptation, and it can significantly improve the result relative to the baseline model by 2.61% higher in Rank-1 metric on Market-1501. In addition, domain adaptation (IIF w/ DA) can further improve the performance by 8.01% mAP on Market-1501, which shows that DA can reduce the impact of domain shift efficiently.

Note that, the method which applies illumination-invariant feature with domain adaptation (IIF w/ DA) is not effective on DukeMTMC-reID because there is no significant change in illumination. Therefore, the proposed IIF only improves the performances slightly on the DukeMTMC-reID.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 5‑5 Ablation studies of hard clustering mining | | | | |
|  | Market-1501 | | DukeMTMC-reID | |
| Method | Rank-1 | mAP | Rank-1 | mAP |
| AWCL | 87.79% | 72.19% | 76.39% | 59.02% |
| AWCL + HCM | 88.65% | 73.05% | 76.70% | 60.12% |
|  |  |  |  |  |

1. *The effect of the hard clustering mining*

In this part, we do the experiment to verify that hard clustering mining can get the highly performance on two public datasets. Table 5‑5 shows the effect of our proposed hard clustering mining. The experimental results illustrate hard clustering mining can boost the performance by 1% higher both in Rank-1 and mAP because model can pay attention on the hard negative sample.

1. *The effect of all proposed method*

Last, we combine all of the proposed method together and see how each method affect our model as shown in Table 5‑6. We observe that without the proposed adaptive weighted clustering loss, our method cannot achieve higher performance than 70% mAP on Market-1501. And the result shows that the proposed adaptive weighted clustering loss can scientifically improve our result on both datasets. After applying the IIF, the mAP improves significantly from 72.19% to 75.02% on Market-1501 and by 1.35% on DukeMTMC-reID. Furthermore, adopting hard clustering mining further improve the performance by 1% on the Market-1501.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 5‑6 Ablation studies of the proposed method | | | | |
|  | Market-1501 | | DukeMTMC | |
| Method | Rank-1 | mAP | Rank-1 | mAP |
| baseline | 81.71% | 60.34% | 71.85% | 52.50% |
| IIF w/ DA | 86.01% | 68.35% | 75.62% | 55.65% |
| AWCL | 87.79% | 72.19% | 76.39% | 59.02% |
| IIF w/ DA + AWCL | 89.57% | 75.02% | 78.84% | 60.37% |
| IIF w/ DA + AWCL + HCM  (ours) | 90.73% | 76.32% | 78.75% | 60.48% |
|  |  |  |  |  |

### The Result of Market-1501 Dataset

To evaluate our proposed method, we follow the evaluation metrics suggested by [61] and report the CMC and mAP accuracy in percentage terms. In the experiment, all 1501 identities are split into training set and testing sets with containing 750 identities and 751 identities. While retrieving stage, the image captured in the same camera with the same identity should to be ignored, which means that if the query image is captured in camera1 and the query identity is ID 1, then we have to ignore the ID 1 that was also captured in camera1 in the gallery.

Table 5‑7 shows the performance of several methods. Except for [61] and [65], all of the methods are learning-based feature. We can see that the hand-craft-based methods such as Bow + kissme [61] and Hybrid [65] achieve low performance because hand-craft feature cannot apply to in different viewpoints. CRAFT-MFA [66] has a relatively high accuracies by learning-based feature with feature fusion. The overall of deep-learning-based methods such as SVDNet [67], PSE [20], CamStyle [68], and AWTL [25] can achieve highly performance because the end-to-end training can easily teach model to extract view-invariant feature and discriminative feature. Table 5‑6 shows that, the proposed adaptive weighted clustering loss (AWCL) outperforms all existing methods on both Rank-1 and mAP. After training on the synthetic dataset to obtain illumination-invariant property with domain adaptation, our method (IIF w/ DA + AWCL + HCM) has achieve 90.73% in Rank-1 evaluation metric, which is about 2.61% better than the CamStyle [68], and 4.56% higher in mAP.

It is worthy to note that our method only use ResNet50 as our backbone, and only depends on RGB information that easily implement on the real-world applications.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 5‑7 Comparison of the state-of-the-art methods on the market-1501 | | | |
| Method | Data type | Rank-1 | mAP |
| Bow + kissme [61] | RGB | 42.64% | 19.74% |
| Hybrid [65] | RGB | 48.15% | 29.94% |
| CRAFT-MFA [66] | RGB | 71.80% | 45.50% |
| SVDNet [67] | RGB | 82.30% | 62.10% |
| PSE [20] | RGB + Skeleton | 87.70% | 69.00% |
| CamStyle [68] | RGB | 88.12% | 68.72% |
| AWTL [25] | RGB | 86.11% | 70.83% |
| AWTL + HNM [25] | RGB | 86.94% | 71.76% |
| IIF w/ DA + AWCL + HCM (ours) | RGB | **90.73%** | **76.32%** |
|  |  |  |  |

### The Result of DukeMTMC-reID Dataset

We follow the evaluation rules on the DukeMTMC-reID which is the same as Market-1501. The quantitative results are showed in Table 5‑8, Our method (IIF w/ DA + AWCL + HCM) outperforms most state-of-the-arts and the performance approach PSE [20] which uses RGB and skeleton as input data. Although our performance is lower than the PSE, our model is more efficient than the PSE because PSE need to perform pose estimation before person re-identification, which is need highly computational cost. The same as the result on the Market-1501, hand-craft-based methods such as Bow + kissme [61] and LOMO + XQDA [69] achieve low performance on the DukeMTMC-reID. Basel [70] is the deep-learning baseline which get better performance compared to the hand-craft based methods. Besides, we found that the performance of the DukeMTMC-reID is lower than the Market-1501 because the viewpoints of the DukeMTMC-reID will change dramatically, which cause the dataset more challenging.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 5‑8 Comparison of the state-of-the-art methods on the DukeMTMC-reID | | | |
| Method | Data type | Rank-1 | mAP |
| Bow + kissme [61] | RGB | 25.13% | 12.17% |
| LOMO + XQDA [69] | RGB | 30.75% | 17.04% |
| Basel [70] | RGB | 65.22% | 44.99% |
| SVDNet [67] | RGB | 76.70% | 56.80% |
| PSE [20] | RGB + Skeleton | **79.80%** | **62.00%** |
| CamStyle [68] | RGB | 78.32% | 57.61% |
| AWTL [25] | RGB | 75.31% | 57.25% |
| AWTL + HNM [25] | RGB | 77.69% | 58.74% |
| IIF w/ DA + AWCL + HCM (ours) | RGB | 78.75% | 60.48% |

### The Result of ACL-reID

We follow the public evaluation metrics which is described in Section 5.3.3, and the experimental results are shown on Table 5‑9. In this table, we can find that with the use of the proposed adaptive weighted clustering, the mAP can increase 7.1%. Furthermore, when applying all of the proposed methods (IIF w/ DA + AWCL + HCM) the results achieve 100% the Rank-1 and 91% in mAP, which represents our model can success fine-tuning to our environment.

|  |  |  |
| --- | --- | --- |
| Table 5‑9 Comparison of the person re-identification on ACL-reID dataset | | |
| Methods | Rank-1 | mAP |
| baseline | 92% | 78.44% |
| AWCL | 96% | 85.54% |
| IIF w/ DA + AWCL | 100% | 90.18% |
| IIF w/ DA + AWCL + HCM | 100% | 91.34% |

# Conclusion and Future Works

In this thesis, we propose a novel cross-illumination person re-identification system which use a deep Convolutional Neural Network (CNN) to extract the robust high-level feature, and the proposed Illumination-Invariant Feature (IIF) can encode different kinds of illumination conditions into the same feature space.

In order to reduce training time, we propose an adaptive weighted clustering loss (AWCL). In comparison with triplet loss () and contrastive loss (), time complexity of the proposed AWCL is equal to . And the performance of the proposed AWCL is better than triplet loss and contrastive loss. The experimental results show that our method (II w/ DA + AWCL + HCM) has achieve 76.32 % in mean average precision on the Market-1501 and 60.48% on the DukeMTMC-reID, which is better than most state-of-the-art methods. Besides, our method is more efficient and time-consuming when training model.

At last, we collect the ACL re-ID dataset to enable our person re-identification model being used in our environment, the dataset contains 40 person identities across 5 cameras in CSIE-Der Tian Hell’s second floor. In addition, we design the 3-stage dataset collection process to automatically cut the video into a person re-identification dataset. To quantify the fine-tuning result on the ACL re-ID, we perform the experiment and the results show that our model can successfully fine-tune our environment.

In future work, we plan to develop a person re-identification system, which contains the human detection module and person re-identification module. Hence, we need to design the suitable human detection model which can detect human in our environment.

1. REFERENCE

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984,* 2016.

[2] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*, 2008, pp. 275: 1-10: British Machine Vision Association.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, no. 2, pp. 91-110, 2004.

[4] 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,* vol. 32, no. 9, pp. 1627-1645, 2010.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision,* vol. 88, no. 2, pp. 303-338, 2010.

[7] T.-W. Hsu, Y.-H. Yang, T.-H. Yeh, A.-S. Liu, L.-C. Fu, and Y.-C. Zeng, "Privacy free indoor action detection system using top-view depth camera based on key-poses," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 004058-004063: IEEE.

[8] T.-E. Tseng, A.-S. Liu, P.-H. Hsiao, C.-M. Huang, and L.-C. Fu, "Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 4077-4082: IEEE.

[9] S.-C. Lin, A.-S. Liu, T.-W. Hsu, and L.-C. Fu, "Representative body points on top-view depth sequences for daily activity recognition," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2968-2973: IEEE.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886--893: IEEE Computer Society.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[14] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263-7271.

[15] 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*, 2015, pp. 91-99.

[16] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335-1344.

[17] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007, vol. 3, no. 5, pp. 1-7: Citeseer.

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

[19] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152-159.

[20] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 420-429.

[21] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1354-1362.

[22] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960-3969.

[23] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062-1071.

[24] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 932-940.

[25] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6036-6046.

[26] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291-7299.

[27] 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*, 2016, pp. 17-35: Springer.

[28] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. Frank Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 172-178.

[29] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244,* 2016.

[30] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1306-1315.

[31] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495,* 2014.

[32] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[33] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 994-1003.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.

[35] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, vol. 2, pp. 1735-1742: IEEE.

[36] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815-823.

[37] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, 2013, pp. 160-172: Springer.

[38] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, vol. 96, no. 34, pp. 226-231.

[39] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 7, pp. 881-892, 2002.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[41] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision,* vol. 115, no. 3, pp. 211-252, 2015.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014, pp. 818-833: Springer.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[44] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks,* vol. 2, no. 5, pp. 359-366, 1989.

[45] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, no. 02, pp. 107-116, 1998.

[46] 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*, 2016, pp. 770-778.

[47] G. Varol *et al.*, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 109-117.

[48] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep cnns in re-identification," *Computer Vision and Image Understanding,* vol. 167, pp. 50-62, 2018.

[49] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 189-205.

[50] 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)*, 2018, pp. 480-496.

[51] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology,* 2018.

[52] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724,* 2017.

[53] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bags of Tricks and A Strong Baseline for Deep Person Re-identification," *arXiv preprint arXiv:1903.07071,* 2019.

[54] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, no. Nov, pp. 2579-2605, 2008.

[55] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167-7176.

[56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, 2016, pp. 499-515: Springer.

[57] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 274-282: ACM.

[58] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," *arXiv preprint arXiv:1708.02551,* 2017.

[59] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690-4699.

[60] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737,* 2017.

[61] 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*, 2015, pp. 1116-1124.

[62] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754-3762.

[63] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315-323.

[64] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*: Springer, 2010, pp. 177-186.

[65] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition,* vol. 65, pp. 238-250, 2017.

[66] 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,* vol. 40, no. 2, pp. 392-408, 2017.

[67] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800-3808.

[68] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157-5166.

[69] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197-2206.

[70] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* vol. 14, no. 1, p. 13, 2018.