Deep Learning Approaches to Person Re-identification

Yutian Lin

Supervisor: Dr. Yi Yang

Centre for Artificial Intelligence
Faculty of Engineering and Information Technology

University of Technology Sydney

This dissertation is submitted for the degree of PhD

August 2019

Declaration

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

August 2019

Acknowledgements

It is a long journey to pursue a PhD. I could not make it without the help from many people. Thanks to all the wonderful people I have worked with during my PhD study.

First and foremost, I would like to thank my advisor Prof. Yi Yang. Pursuing a PhD degree under his supervision is the most important and lucky decision I have ever made. In the past three years, I learned scientific thoughts and very useful technical skills from him. The insistence and enthusiasm he has for research is contagious and motivational for me in the PhD pursuit. I appreciate all his contributions of time, ideas, and funding during my doctoral studies. Yi also gave me many chance to visit other research groups that I can have the opportunity to cooperate with different great people. Thanks a lot for letting me visit Carnegie Mellon University, Texas State University and have a chance for the internship.

I would like to thank Liang Zheng, the past postdoc at the University of Technology Sydney, who helps me a lot in the first two years. He introduced me to the person re-identification area and provided many ideas to my papers. Liang was always patient to help me, communicated with me and work until late night to polish my papers. Thanks a lot for his dedication in the early period of my doctoral studies.

I would like to thank Alexander Hauptmann, my host professor at Carnegie Mellon University. The communications with Alex's group inspired me a lot to my research.

I would like to thank Prof. Yan Yan, my host professor at Texas State University. I'm grateful for his patient and help in both research and life.

The researches are partially supported by Data to Decisions CRC. Thanks to Data to Decisions CRC for the scholarship support.

Thanks to the members of Yi's group, who have contributed immensely to my personal and professional time at UTS. The group has been a source of friendships as well as good advice and collaboration. Thanks to Zhongwen Xu, Xiaojun Chang, Yan Yan, Guoliang Kang, Wenhe Liu, Linchao Zhu, Pingbo Pan, Zhedong Zheng, Hehe Fan, Xuanyi Dong, Yanbin Liu, Peike Li, Qianyu Feng, Jiaxu Miao, Xiaohan Wang, Hu Zhang, Guang Li, Ruoyu Liu, Zhun Zhong, Yawei Luo, Qingji Guan. Special thanks

to Zhedong Zheng, who help me a lot in the early years of my research. I learned a lot in the seminar he held.

Thanks to my parents Jianqiang Lin and Ling Gao for all their love, selfless support and encouragement. They worked and sacrificed to give me every opportunity to succeed and they are always there when I need help.

Thank my beloved boyfriend Yu. He has been my best friend, classmate, and collaborator from our high school to the doctoral study. There are so many days we stay up late together for papers and researches. He always helps me and encourages me when I face with troubles in my work or life. I could not complete my PhD study without his tremendous support and deep love.

Abstract

Recent years witnessed a dramatic increase of the surveillance cameras in the city. There is thus an urgent demand for person re-identification (re-ID) algorithms. Person re-identification aims to find the target person in other non-overlapping camera views, which is critical in practical applications. In this thesis, I present my research on person re-ID in three settings: supervised re-ID, one-example re-ID and unsupervised re-ID. For supervised setting, a re-ranking algorithm is introduced that can improve the existing re-ID results with Bayesian query expansion. We also investigate pedestrian attributes for re-ID that learns a re-ID embedding and at the same time predicts pedestrian attributes. Since supervised methods require a large amount of annotated training data, which is expensive and not applicable for real-world applications, two re-ID methods on the one-example setting are studied. We also propose an unsupervised re-ID method that jointly optimizes a CNN model and the relationship among the individual samples. The experimental results demonstrate that our algorithm is not only superior to state-of-the-art unsupervised re-ID approaches but also performs favourably than competing transfer learning and semi-supervised learning methods. Finally, I make conclusions on my work and put forward some future directions on the re-ID task.

Table of contents

Chapter 1

Introduction

1.1 Background

Person re-identification (re-ID) is a fundamental task in video surveillance and has become a focus in the recent vision research. Given an image/video of a person, the goal of person re-ID is to find the aimed person in other non-overlapping camera views. Person re-ID is challenging, partially due to the distinct characteristics of different cameras, such as view points, illuminations, *etc.*

1.1.1 Supervised Person Re-identification

There are numerous supervised re-ID methods that achieve impressive performances. These methods mainly focus on designing feature representations [140] or learning robust distance metrics [54, 145]. Recently, deep learning methods achieve great success [51, 100, 142, 146] by simultaneously learning the image representations and similarities.

To make further progress based on the existing re-ID methods, we consider a query expansion method that post-process the initial ranking list to get more precise performance. In Chapter 3, we propose a Bayesian Query Expansion method, that a Bayesian model is used to predict true matches in the gallery, which are used to fuse a new query.

Since person re-ID [59, 83, 111, 156] and attribute recognition [1, 12, 153] both imply critical applications in surveillance. We also tried to make use of the detailed attribute information in the re-ID framework. In Chapter 4, we propose a network that learns a re-ID embedding and predicts the pedestrian attributes simultaneously. We show that attributes help improve person re-identification performance.

1.1.2 One-example Person Re-identification

As we illustrated in Section 1.1, most proposed approaches rely on the fully annotated data, *i.e.*, the identity labels of all the tracklets from multiple cross-view cameras. However, it is impractical to annotate very large-scale surveillance videos due to the dramatically increasing cost. Therefore, one-example methods [66, 129] are of particular interest.

The key challenge for the one-shot video-based person re-ID is the label estimation for the abundant unlabeled tracklets [16, 129]. A typical approach is to generate the pseudo labels for the unlabeled data at first. The initial labeled data and some selected pseudo-labeled data are considered as an enlarged training set. Lastly, this new training set is adopted to train the re-ID model.

Most existing methods employ a static strategy to determine the quantity of selected pseudo-labeled data for further training. For example, Fan *et al.*. [16] and Ye *et al.*. [129] compare the prediction confidences of pseudo-labeled samples with a *pre-defined* threshold. The samples with higher confidence over the fixed threshold are then selected for the subsequent training. During iterations, these algorithms select a fixed and large number of pseudo-labeled data from beginning to end. However, it is inappropriate to keep the threshold fixed in the one-shot setting. In this case, the initial model may be not robust due to the very few training samples. Only a few of pseudo-label predictions are reliable and accurate at the initial stage. If one still selects the same number of data as that in the later stages, it will inevitably involve many unreliable predictions. Updating the model with excessive not-yet-reliable data would hinder the subsequent improvement of the model.

In Chapter 5, to better exploit the unlabeled data in one-shot video-based person re-ID, we propose the step-wise learning method **EUG** (**E**xploit the **U**nknown **G**radually), that iteratively updates the CNN by label estimation and model updates. In Chapter 6, we further improve the EUG method by making use of the unlabeled data. Our proposed methods achieve state-of-the-art performance on four large-scale image-based and video-based datasets.

1.1.3 Unsupervised Person Re-identification

Since one-example methods still have limitations that require one sample annotated for an identity. The limited generalization ability motivates the research into unsupervised approaches for person re-ID.

Traditional unsupervised methods focus on hand-crafted features [17, 54, 57], salience analysis [103, 139] and dictionary learning [39]. These methods produce much lower performance than supervised methods and are not applicable to large-scale real-world data. In recent years, some transfer learning methods [11, 28, 80] are proposed upon the success of deep learning [53]. These methods usually learn an identity-discriminative feature embedding on the source dataset, and transfer the learned features to the unseen target domain. However, these methods require a large amount of annotated source data, which cannot be regarded as pure unsupervised approaches.

Previous deep learning based "unsupervised" person re-ID approaches leverage the prior knowledge learned from other re-ID datasets. However, we aim to solve the problem in a more challenge and practical setting, *i.e.*, without any re-ID annotation. To learn discriminate features in this difficult condition, in Chapter 7, we propose a novel Bottom-Up Clustering method (**BUC**) for unsupervised re-ID that maximizes the diversity over the identities while maintaining the similarity within each identity. We conduct extensive experiments on the large-scale image and video re-ID datasets, including Market-1501, DukeMTMC-reID, MARS and DukeMTMC-VideoReID. The experimental results demonstrate that our algorithm is not only superior to state-of-the-art unsupervised re-ID approaches, but also performs favorably than competing transfer learning and semi-supervised learning methods.

1.2 Thesis Organization

This thesis is organized as follows:

- *Chapter 2*: This chapter presents a survey of various methods for person re-identification, including the supervised methods, weakly-supervised methods and unsupervised methods.

- *Chapter 3*: In this chapter, we introduce a supervised Bayesian query expansion method for person re-identification. Given a rank list, a Bayesian model is proposed to calculate the weight for the reliable retrieved images. We then apply weighted average pooling to the feature of the reliable images for query expansion. This method achieves consistent improvement over various baselines.

- *Chapter 4*: In this chapter, we introduce a supervised attribute based method for person re-identification. The method takes advantage of the mutual benefit of the global feature and the attribute information and improve the performance of both of the person re-ID and attribute prediction task.

- *Chapter 5*: In this chapter, we focus on one-example person re-identification. For each identity, there is only one sample is annotated with a label. We use these labeled samples to train a network for initialization and then gradually exploit unlabeled samples with a reliable pseudo label for further training.

- *Chapter 6*: We make progress upon the method introduced in chapter 5. We introduce a one-example method suitable for both video-based and image-based re-ID. In addition to gradually exploiting the reliable unlabeled samples, we further learn from the unselected unlabeled samples and achieve better performance.

- *Chapter 7*: In this chapter, we introduce an unsupervised person re-ID method that iteratively maximizes the diversity over the identities while maintaining the similarity within each identity. Without any annotation, this method achieves better performance than the one-example or transfer learning methods.

- *Chapter 8*: A brief summary of the thesis contents and its contributions are given in the final chapter. Recommendation for future works is given as well.

Chapter 2

Literature Review

2.1 Supervised Person Re-identification

Hand-crafted person re-ID. To address the re-ID problem, a number of approaches focus on developing robust features. In these studies, many hand-crafted features have been developed, such as color and texture histograms [17, 23, 54]. Zhao *et al.* [138, 140] propose a feature that combined SIFT feature with color histogram. In [98] local feature is combined with texture, and in [109, 143] local feature is combined with SIFT based on the bag-of-word (BoW) structure. Liao *et al.* propose Local Maximal Occurrance (LOMO) descriptor, which analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation. In [74], Matsukawa *et al.* propose a hierarchical Gaussian descriptor, which calculates both mean and covariance information of pixel features in each patch and region hierarchy.

CNN-based person re-ID. CNN-based methods are dominating the re-ID community upon the success of deep learning [27, 34, 100, 146, 151].A branch of works learning deep metrics [13, 51, 71] that image pairs or triplets are fed into the network. Usually, the spatial constraints are integrated into the similarity learning process [2, 51]. For example, in [100], a gating function is inserted in each convolutional layer, so that some subtle difference between two input images can be captured. Generally speaking, deep metric learning methods have advantages in training on relatively small datasets, but its efficiency on larger galleries may be compromised. Another branch of works learning deep representations [112, 116, 146, 151]. Xiao *et al.* [116] propose to learn a generic feature embedding by training a classification model from multiple domains with a domain guided dropout. In [146], the combination of verification and classification losses is proven effective. Xu *et al.* [120] propose a Pose guided Part Attention (PPA)is learned to extract attention-aware feature for body parts from a base network. Then the features of body parts are further re-weighted, resulting in the final feature vector. Since GAN proposed by Goodfellow *et al.* [22], methods utilizing GAN [107, 147] have been proposed to tackle re-ID. In [107], a Person Transfer Generative Adversarial Network (PTGAN) is proposed to transfer the image style from one dataset to another while keeping the identity information to bridge the domain gap. In [73], a dictionary-learning scheme is applied to transfer the feature learned by object recognition and person detection (source domains) to person re-ID (target domain)

2.2 Weakly Supervised Person Re-identification

Semi-supervised learning methods. Semi-supervised setting assume that few data from the training set are labeled data and rest of the training set are used as unlabeled data. In [49], a novel semi-supervised region metric learning method is proposed, that estimates positive neighbors to generate positive regions and learn a discriminative region-to-point metric. Dictionary learning is originally designed for unsupervised learning and is adopted for semi-supervised re-ID. In [61], two coupled dictionaries that relate to the gallery and probe cameras are jointly learned in the training phase from both labeled and unlabeled images. Other methods [3, 49] utilize another dataset to pretrain the model, and then apply it to the target dataset. In [3], a one-example learning approach is proposed, where one part (texture) of the metric is transferred directly to the target dataset. The second part (color) is learned using patch-based metric learning.

There are some works that focus on the few-example video-based re-ID task. Ye *et al.* [129] propose a dynamic graph matching (DGM) method, which iteratively updates the image graph and the label estimation to learn a better feature space with intermediate estimated labels. Liu *et al.* [66] update the classifier with K-reciprocal Nearest Neighbors (KNN) in the gallery set, and refine the nearest neighbors by apply negative sample mining with KNN in the query set. Even though [66, 129] claim that they are *unsupervised* methods, they are *one-example* methods in experiments, because both of them require at least one labeled tracklet for each identity.

Unsupervised domain adaptation. Recently, cross-domain transfer learning is adopted in the unsupervised re-ID task [28, 105], where information from an external source dataset is utilized. In [67], a classifier trained on source dataset is applied to the unlabeled target dataset to learn the pedestrians' spatial-temporal patterns, which are further combined with visual features to achieve an improved classifier. Finally, we propose a learning-to-rank based mutual promotion procedure to incrementally optimize the classifiers based on the unlabeled data in the target domain.

Wang *et al.* [105] propose to learn an attribute-semantic and identity discriminative representation from the source dataset, which is transferable to the target domain. Some methods [11, 108] proposed to learn a similarity preserving generative adversarial network based on CycleGAN [154] to translate images from the source domain to the target domain. In this way, high-quality person images are generated, and person identities are kept and the styles are effectively transformed. The translated images are utilized to train re-ID models in a supervised manner. These methods assume that the label of the source domain is available and apply the learned discriminative model to the target domain. In this work, we propose a fully unsupervised re-ID framework that gradually exploits the similarity within each identity.

2.3 Unsupervised Person Re-identification

The existing fully unsupervised methods usually fall into three categories, designing hand-craft features [17, 54, 57], exploiting localized salience statistics [103, 139] or dictionary learning based methods [39, 123]. However, it is a challenging task to design suitable features for images captured by different cameras, under different illumination and view condition. In [132], camera information is used to learn view-specific projection for each camera view by jointly learning the asymmetric metric and seeking optimal cluster separations. Recently, Lin *et al.* [55] propose a bottom-up clustering framework that jointly optimize a convolutional neural network (CNN) and the relationship among the individual samples.

There are also some recent works [66, 127, 129] focusing on the unsupervised video-based re-ID. However, these methods require some very useful annotations of the dataset, *i.e.*, the total number of identities and their appearance. To conduct experiments, they annotate each identity with a labeled video tracklet, which only reduces part of the annotation workload. As discussed in [114], these approaches are actually the one-example methods.

2.4 Datasets

The Market-1501 dataset [142] is one of the largest person re-ID dataset, which contains 32,668 gallery images, 3,368 query images captured by 6 cameras. It also includes 500K irrelevant images as distractor set, which makes the dataset even more challenging. This dataset is captured in a noisy campus environment. It is very challenging due to the illumination and viewpoint changes across the cameras, and occlusions in the views. Following the experimental protocol in [142], the dataset is split into 751 identities for training and 750 identities for testing.

The DukeMTMC-reID dataset [147] is a subset of the DukeMTMC dataset [84]. It contains 1,812 identities captured by 8 cameras. A number of 1,404 identities appear in more than two cameras, and the rest 408 IDs are distractor images. Using the evaluation protocol specified in [147], the training and testing set both contain 702 IDs, with 16,522 training images and 17,661 gallery images, respectively.

The DukeMTMC-VideoReID dataset is a subset of DukeMTMC[84], which is specially for video-based re-ID. Since this dataset is manual annotated, each identity only has one tracklet under a camera. The pedestrian images are cropped from the videos for 12 frames every second to generate a tracklet. The dataset is split following the protocol in [147], *i.e.*, 702 identities for training, 702 identities for testing, and 408 identities as the distractors. Totally, 369,656 frames of 2,196 tracklets are generated for training, and 445,764 frames of 2,636 tracklets are generated for testing and distractors.

The MARS dataset [141] is the largest video re-ID dataset, which contains 1261 individuals and around 20,000 video sequences captured by six cameras. The MARS dataset is divided into train and test sets, containing 631 and 630 identities respectively. Each identity has 13.2 tracklets on average. The dataset is captured in a noisy campus environment, thus suffering from significant viewpoint changes, pose variation, and illumination changes. For the MARS dataset, we use the multi-shot protocol [141] in our experiment.

The CUHK03 dataset[51] contains 13,164 images of 1,360 pedestrians captured by six cameras. Each identity appears in two disjoint camera views. This dataset is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, and occlusions. Note that the original evaluation protocol of CUHK03 has 20 train/test splits. To maintain consistency with other datasets, we use the train/test protocol proposed in [149]: 7,365 images of 767 identities are used as the training set. 5,332 images and 1,400 images of the remaining 700 identities are used as the gallery set and the query set, respectively. We only conduct experiments on the DPM-detected images.

The PETA dataset [12] is a large person attribute recognition dataset that annotated with 61 binary attributes and 4 multi-class attributes for 19,000 images. Following [12], 35 most important and interesting attributes are used in our experiments. To investigate the complementarity of re-ID and attribute recognition, we re-split this dataset for re-ID task. We randomly select 9,500 images of 4,558 identities for training, and 7,600 images for testing. As a result, we generate 423 query images and 7,177 gallery images.

Chapter 3

Query Expansion Method to Person Re-identification

In this chapter, we investigate the query expansion method to person re-identification (re-ID). Person re-identification is challenging because pedestrians may exhibit distinct appearance under different cameras. Given a query image, previous methods usually output the person retrieval results directly, which may perform badly due to the limited information provided by the single query image. To mine more query information, we add an expansion step to post-process the initial ranking list. The intuition is that a true match in the gallery may be difficult to be found by the query alone, but it can be easily retrieved by other true matches in the initial ranking list. In this chapter, we propose the Bayesian Query Expansion (BQE) method to generate a new query with information from the initial ranking list. The Bayesian model is used to predict true matches in the gallery. We apply pooling on the features of these "true matches" to get a single vector, *i.e.*, the expanded new query, with which the retrieval process is performed again to obtain the final results. We evaluate BQE with various feature extraction methods and distance metric learning methods on four large-scale re-ID datasets. We observe consistent improvement over all the baselines and report competitive performances compared with the state-of-the-art results.

3.1 Introduction

In a different view from previous works on the feature or metric learning [17, 116], we aim to improve the existing re-ID algorithms by query expansion. We investigate some large-scale multi-camera datasets, such as Market-1501 [142], and find that the same identity's appearance in different cameras usually undergoes large variance. On

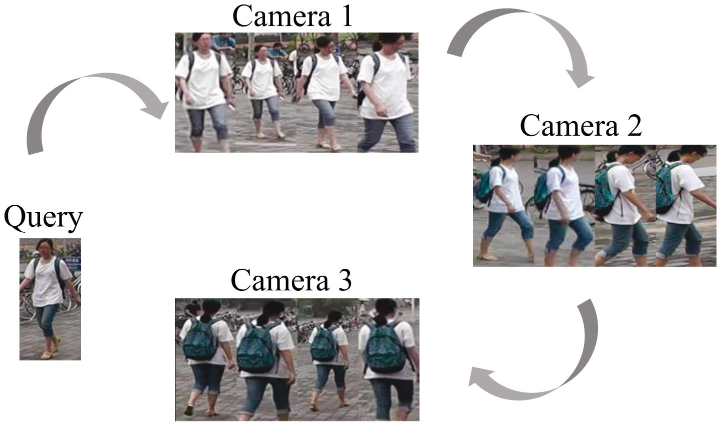


Fig. 3.1 An example of how query expansion works. Given a query image, true matches in camera 1 somehow can be easily found. With the additional discriminative cues (the dark green bag) of retrieved images in camera 1, true matches in camera 2 can be found. Similarly, we can then obtain the true matches in camera 3.

the contrary, the appearance of each pedestrian under the same camera is usually similar. As shown in Fig. 3.1, query expansion can help to improve the original re-ID performance.

Inspired by the above considerations, we propose a Bayesian query expansion re-ranking algorithm to improve the performance of the existing re-ID methods. In a nutshell, after investigating the initial rank list, a new query is constructed based on the top returned images. Since the top-ranked images may be false matches, we develop a Bayesian framework to identify if a returned candidate is a true match; only the candidates with high probability scores will be used for query expansion. In this manner, the expanded query feature will contain more discriminative cues that may be absent in the initial query and will be used to search the system for a second time to improve re-ID recall. The pipeline of our method is shown in Fig. 3.2.

To evaluate the performance of the proposed method, we perform experiments on four large-scale datasets including Market-1501 [142], DukeMTMC-reID [147], MARS [141], and CUHK03 [51]. We show that BQE effectively improves the performance of the baseline systems, and that it has comparable accuracy with several competing re-ranking methods while enjoying efficient offline computations and robustness of parameter changes. Using the expanded queries, we are able to achieve very competitive results to the state-of-the-art methods.

Our contributions are summarized as follows:

(1) We propose a Bayesian Query Expansion (BQE) method for re-ID re-ranking. BQE generates a new query with information from the initial gallery ranking list, which is used to retrieve the gallery images again.

(2) The Bayesian model is trained by the distances between images in the training data and is used to calculate the probability of images in the gallery being a "true match". We apply pooling on the features of the "true matches" to get a vector for query expansion.

(3) Our method effectively improves the person re-ID performance on several datasets, including Market-1501, CUHK03, DukeMTMC-reID and MARS. We also achieve the state-of-the-art accuracy on Market1501.

3.2 Related Work

Re-ranking methods in multimedia retrieval. Re-ranking methods have widely applied in multimedia retrieval. Some of the algorithms are with human interaction [85, 122], others are implemented without any extra information [26, 63, 148]. For the second type, the re-ranking methods highly depend on the query and the initial ranking list [77].

Many works of query expansion are proposed in the field of image retrieval [9, 10] or text retrieval [102, 119]. In such works, the top-ranked images or documents from the original rank list are used to generate a new query that can be used to obtain a new ranking list. Nevertheless, the effects of all query expansion methods are highly affected by the initial matching results.

Another re-ranking method Pseudo Relevance Feedback(PRF) [63, 64, 126] shows a similar motivation, which assumes the top ranked sample as "pseudo relevant" to address the re-ranking problem. In [33], Jain *et al.* use pseudo relevant in learning method to classify the remaining samples into relevant or irrelevant classes. In [124], pseudo relevant is used as extent query to retrieve the ranking list. Lee *et al.* [46] propose a cluster-based re-sampling method to select better pseudo-relevant documents based on the relevance model. In [125], the lowest ranked images are used as negative examples, and the initial query is used as a positive example to train classifiers. Finally the images are ranked by the confidence scores. In [64], a set of positive pairs in the initial ranking list are selected and used to learns a re-ranking model by Ranking SVM.

Re-ranking methods in re-ID. Comparing with retrieval, fewer studies have investigated the re-ranking method in the field of person re-ID, some of which are presented with human in the loop and others fully automatic.

Liu *et al.*[58] present a post-rank optimization (POP) model which constructs an incremental affinity graph and the negative selections are propagated to their neighbors to refine their search. Further, Wang *et al.* [104] propose a Human Verification Incremental Learning model (HVIL), which constructs an incrementally optimized ranking function and is updated in real-time.

Other researches pay attention to automatic re-ranking method. Ma *et al.* [68] learn an adaptive function specific for each query, which combines the base score function and query match estimated by modeling query variations. In [20], the content and context information is analyzed. And then the visual ambiguities common are removed to re-rank the initial ranking list. Leng *et al.* [47] calculate the matching rates of common k-nearest neighbors between every two bidirectional ranking lists as context and content similarity, and the rates are used to revise the initial query result. Li *et al.* [50] propose a common Near-Neighbor analysis, which analyzes the pair of samples of neighbors using both relative and direct information for re-ranking. In [128], the similarity of top retrieved images and last retrieved images are calculated by different baseline methods, then similarity ranking aggregation and dissimilarity ranking aggregation are used to optimize the ranking result. In [149], Zhong *et al.* encode the top-k retrieved images as the k-reciprocal feature and use it for re-ranking under the Jaccard distance.

3.3 Proposed Method

3.3.1 Problem Formulation

In this section, we present the Bayesian query expansion framework. The overview pipeline of our method is shown in Fig. 3.2. Briefly, our system consists of a Bayesian model (see Section 3.3.2) and the query expansion process (see Section 3.3.3).

In more detail, the dataset is divided into three parts: query, gallery and training data. During the offline procedure, a Bayesian posterior estimator is firstly trained on the training data. Given a distance metric, the Bayesian model can predict the probability of a candidate being a true match. During online retrieval, an initial ranking list is obtained after computing the similarities between the query and the gallery images. Based on the ranking list, the probability of each candidate being a true match is computed by the Bayesian model. Then, features of images in the initial ranking list with high probability are pooled together with the original query, such that a new query is issued to perform another round of retrieval. Note that, after the

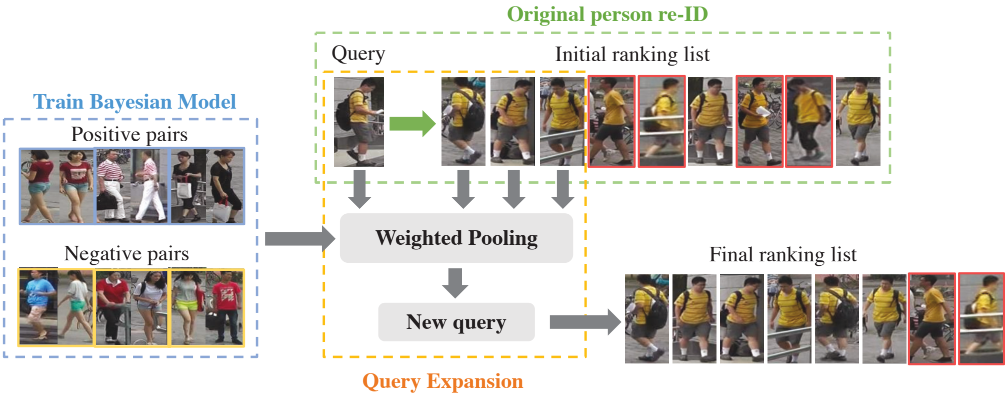


Fig. 3.2 Overview of the proposed approach for person re-identification. The Bayesian model is trained on the distances of relevant image pairs and irrelevant image pairs. Given a query, an initial ranking list is first obtained by a normal retrieval method. The features of top-ranked candidates are then pooled with weights calculated by Bayesian model to generate a new query. Finally, the new query is used for query expansion. The red bounding boxes denote the false matches in the ranking list.

new round of retrieval, the query expansion process can be leveraged again, resulting in an iterative algorithm.

Formally, each image is represented by a d-dimensional feature vector, denoted by $x \in \mathbb{R}^d$. Let $T = \{x_i^t | i = 1, 2, ...M\}$ be the training set, and $G = \{x_i^g | i = 1, 2, ...N\}$ the gallery set. Then the identity label of training image x_i^t, query image q and gallery image x_i^g are denoted as l_i^t, l^q and l_i^g. Let $d(q, x_i^g)$ denote the distance between a query image q and a gallery image x_i^g. Then the initial ranking list is denoted as $R(q) = [x_1^g, x_2^g, ...x_n^g]$, where $d(q, x_i^g) < d(q, x_{i+1}^g)$. Our goal is to re-rank the initial ranking list based on an offline-trained Bayesian model.

3.3.2 Bayesian Model

In spirit, the Bayesian model characterizes the matching score distribution of the true matches and the false matches. The model is created on the training set and deployed during testing to estimate the probability of a top-ranked image being a true match to the query.

Let us formulate the person re-ID re-ranking problem in a more formal way. For each image x in the ranking list returned from a person re-ID system, there is a distance computed by the learned metric. Since images with small distance to the query will be listed at the top, we wonder if we can re-rank the image list using the top images to improve the performance. Selecting the candidate images is crucial, because false matches will have opposite effect on the performance. The problem is, can we estimate the probability of an image being a true match when given a distance between the query and the image, *i.e.*, P(x is a true match | d(x,q)) ? As shown in Fig. 3.3, images of the same or different identities usually have obvious different range of distance, so that the distance can help us distinguish the candidates. In this chapter, the Bayesian model is used to estimate the probability of relevance of an image in the ranking list.

To be specific, for a query q and a gallery image x_i^g, we propose to compute the probability that the two images belong to the same identity given the distance between the two images, *i.e.*, $P(l^q = l_i^g | d(q, x_i^g))$.

By Bayes' theorem, the probability can be rewritten as follows,

$$P(l^q = l_i^g | d(q, x_i^g)) = \frac{P(d(q, x_i^g) | l^q = l_i^g) P(l^q = l_i^g)}{P(d(q, x_i^g))}, \tag{3.1}$$

where $P(d(q, x_i^g))$ can be calculated by,

$$\begin{aligned} P(d(q, x_i^g)) = {} & P(d(q, x_i^g) | l^q = l_i^g) P(l^q = l_i^g) \\ & + P(d(q, x_i^g) | l^q \neq l_i^g) P(l^q \neq l_i^g). \end{aligned} \tag{3.2}$$

We can estimate the probability using the training data. Here, we directly use $P(l_i^t = l_j^t)$ and $P(l_i^t \neq l_j^t)$ to approximate the value of $P(l^q = l_i^g)$ and $P(l^q \neq l_i^g)$. To calculate $P(d(q, x_i^g) | l^q = l_i^g)$, and $P(d(q, x_i^g) | l^q \neq l_i^g)$, we calculate the distances between every images in the training data, and use a range of distance to replace the exact value of distance. We divide the range of distance ($d(q, x_i^g)$ values) into M intervals. Then the number of candidates within each interval is counted. As shown in Fig. 3.3, each bar represents a interval. Assume that $d(q, x_i^g)$ falls in $[0.2, 0.3]$, then $P(d(q, x_i^g) | l^q = l_i^g)$ can be calculated by the number of candidates divided by the frequency of the red bar in this interval. $P(d(q, x_i^g) | l^q \neq l_i^g)$ can be calculated in a similar way. The number of intervals M is chosen based on the size of the dataset. Note that if a distance in the test phase is larger than the upper bound (or smaller than the lower bound) in the training phase, we use the result of the upper (or lower bound).

3.3.3 Query Expansion

For query expansion, a new query is issued to re-rank the candidates. Here only K candidates with high probability are pooled as a new query, where $K \leq$ *the number of true matches* and $K \ll n$. The value of K is evaluated in Section 3.4.3.

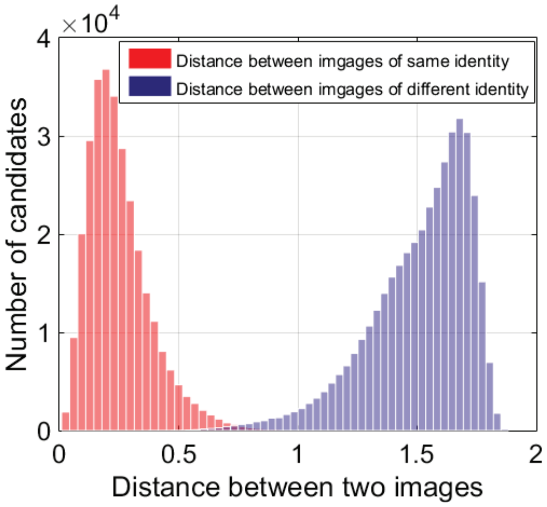


Fig. 3.3 Histogram of Euclidean distance on the training set of the Market-1501 dataset. The red bars represent the distance between the images of the same identity, while the blue bars represent the distance between images of different identities.

There are two simple strategies for query expansion: average query expansion (AQE) and max query expansion (MQE). For these two methods, average pooling and max pooling are used to fuse the feature of the query image and the top ranked candidates, respectively. For AQE, the expanded query is calculated as:

$$q_{new} = \frac{\sum_{i=1}^{K} x_i^g + q}{K+1}. \tag{3.3}$$

The shortage of these strategies is that the effectiveness strongly relies on the quality of the initial ranking list and the value of parameter K. When the initial ranking list is not satisfying or the K is large, false matches will be used to construct the new query, which would affect the precision.

To overcome the shortage, we assign different weight to each candidate when doing feature pooling. Given a query, the probability of each candidate being a true match of the query is computed in Section 3.3.2. Then the expanded probe q_{new} of the initial query q is computed by pooling the top K images and query q with the probabilities. Here we simply use average pooling with weight, where the weight is exactly the probability. The formula is as follows:

$$q_{new} = \frac{\sum_{i=1}^{K} P(l^q = l_i^g) * x_i^g + q}{\sum_{i=1}^{K} P(l^q = l_i^g) + 1} \tag{3.4}$$

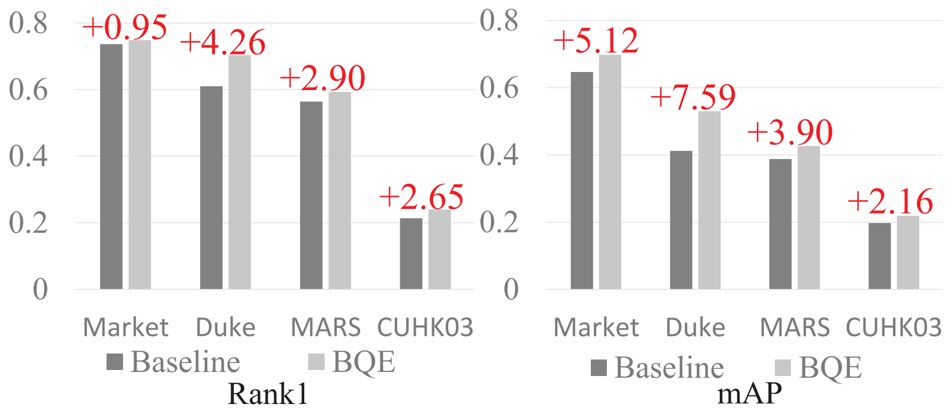


Fig. 3.4 Comparison between the baselines and BQE on four datasets. All the experiments are implemented using IDE and Euclidean distance.

Finally this new query is used to calculate the distance and re-rank the initial ranking list.

More iterations. We assume that the expanded query will lead to a better ranking list, which can produce a better query. Thus we can conduct the procedure of producing ranking list, feature pooling, and query expansion repeatedly. By repeating BQE, the effect will be strengthened. We denote T as the number of iterations, and this parameter is evaluated in Section 3.4.3

3.3.4 Complexity Analysis

A large proportion of the computational cost consists in training Bayesian model and query expansion. Suppose the size of the training set and the gallery set is M and N, respectively. The Bayesian model is computed offline with complexity $\mathcal{O}(M^2)$. For query expansion procedure, we need to compute the probability and construct the new query. The time complexity is $\mathcal{O}(K)$, where K is the number of pooled images. Since parameter $K <$ number of true matches, and $K \ll N$, the complexity can be constrained to $\mathcal{O}(1)$. Then the pairwise distance is computed with complexity $\mathcal{O}(N)$. As a result, for one query, the computation complexity is $\mathcal{O}(N)$.

3.4 Experiments

3.4.1 Implementation Details.

To demonstrate the robustness of this method, we adopt various feature extraction methods and metric learning approaches as baselines.

- ID-discriminative Embedding (IDE) [144]. The IDE extractor is trained on classification model using ResNet-50. It generate a 2048-dim vector for each image.
- Attribute-Person Re-identification Embedding (APR) [56]. The descriptor is trained on classification model using both ID and attribute labels. The network is also trained on ResNet-50 model. For each image, a 2048-dim vector is extracted.
- Bag-of-Words (BoW) descriptor [142]. In the BoW model, local features are aggregated into a global feature vector. For each image, a 5600-dim descriptor is computed.

We also employ three distance learning methods, including Euclidean distance, KISSME and XQDA.

For all the experiments, we set the parameter $K = 5$ on Market-1501, DukeMTMC-reID and MARS. We set $K = 3$ on CUHK03, because the true matches of this dataset are fewer. For all the four datasets, the number of intervals M is set to be 100. When conduct BQE for iterations, the parameter T is set to be 3.

3.4.2 Evaluation of BQE

Comparison with the baselines.

We evaluate if the BQE method outperforms the baselines. Results on the four datasets, *i.e.,* Market-1501, Duke, MARS and CUHK03, are reported in Table 3.1, Table 3.2, Table 3.3 and Table 3.4, respectively. Especially we show the baseline and BQE results on the four datasets using the same setting in Fig. 3.4. We have several observations from these results.

First, our method exceeds the baselines on four datasets. On the Market-1501 dataset, our method consistently improves the rank-1 accuracy and mAP with all features(IDE, APR and Bow model). For example, when using IDE [56] and Euclidean distance, the improvements are +0.95% in rank-1 accuracy and +5.12% in mAP. Moreover, experiments conducted with three metrics all show better results than the

Table 3.1 BQE results on Market-1501. The best and second highest results are highlighted in blue and red.

Methods	rank-1	rank-5	rank-20	mAP
DADM [92]	39.4	-	-	19.6
MBC [99]	45.56	67	82	26.11
SML [36]	45.16	68.12	84	-
DLDA [111]	48.15	-	-	29.94
SL [6]	51.9	-	-	26.35
DNS [133]	55.43	-	-	29.87
LSTM [101]	61.6	-	-	35.3
S-CNN [100]	65.88	-	-	39.55
2Stream [146]*	79.51	90.91	96.23	59.87
GAN [147]*	79.33	-	-	55.95
APR+EU	84.29	93.20	97.00	64.67
APR+EU+BQE	85.24	93.46	97.32	69.79
APR+EU+BQEI	84.91	93.47	96.07	70.74
APR+Kissme	83.90	93.14	97.00	63.34
APR+Kissme+BQE	84.89	93.25	97.35	68.60
APR+XQDA	82.27	92.40	96.80	63.05
APR+XQDA+BQE	83.40	92.71	97.43	67.06
IDE+EU	73.69	88.15	94.83	51.48
IDE+EU+BQE	74.88	88.19	93.37	56.91
IDE+XQDA	72.35	86.78	94.32	50.19
IDE+XQDA+BQE	73.42	85.71	94.49	54.08
IDE+kissme	73.49	88.07	95.25	50.85
IDE+kissme+BQE	74.56	87.52	94.46	55.03
Bow+EU	34.03	50.80	65.91	13.15
Bow+EU+BQE	34.14	50.82	65.93	13.36
Bow+XQDA	41.93	63.26	79.87	21.90
Bow+XQDA+BQE	42.97	63.58	81.02	23.93
Bow+Kissme	41.26	60.38	78.33	20.74
Bow+Kissme+BQE	42.55	60.73	81.22	22.39

baselines, which demonstrate the effectiveness of our method on different distance metrics.

On the DukeMTMC-reID dataset, we test BQE with IDE, APR and three matrix learning methods. In all the settings, the performance of BQE is superior to the baselines. For example, when using APR[56] and Euclidean distance, we observe the most obvious improvement. The performance gain is +4.26% in rank-1 accuracy and +7.59% in mAP.

Consistent findings also hold for the MARS dataset and the CUHK03 dataset, *i.e.*, when IDE and three matrix learning methods are used, BQE improves both rank-1 accuracy and mAP over the baselines.

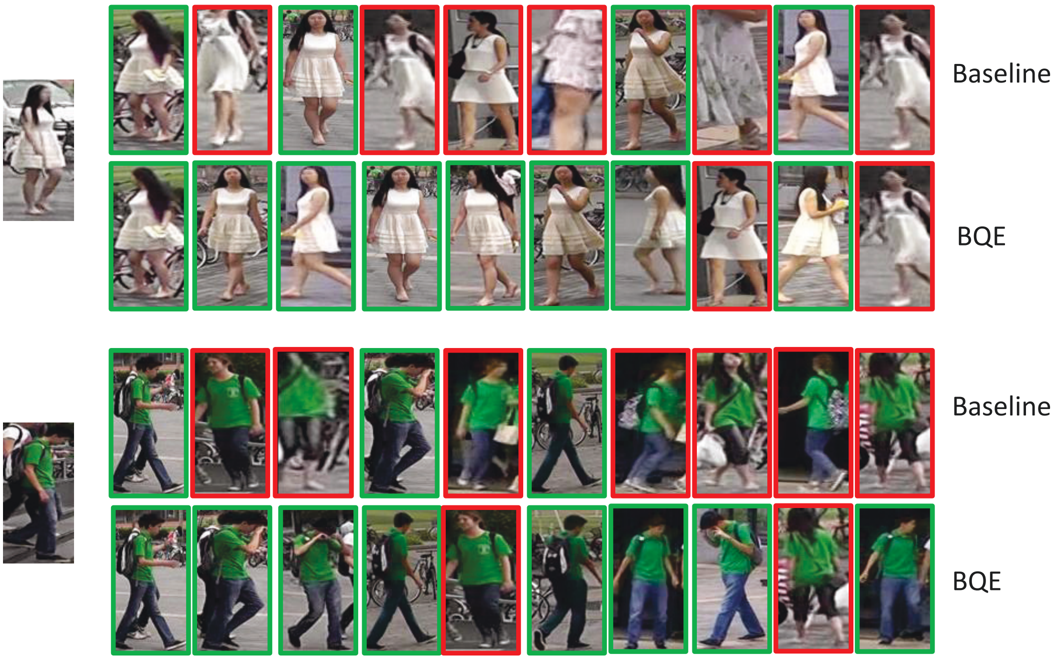


Fig. 3.5 Sample results on the Market-1501 dataset. The green bounding boxes and red bounding boxes denote the true matches and the false ones, respectively.

Second, the improvement of mAP is larger than that of rank-1 accuracy. For example, on the Market-1501 dataset, when using CNN features ([144] or [56]), mAP increases 4% $\sim$ 5%, while rank-1 increases 1% $\sim$ 2%. The results on the DukeMTMC-reID dataset are similar, *i.e.*, the improvement of mAP and rank-1 accuracy is about 5% $\sim$ 11% and 3% $\sim$ 9%, respectively. We speculate that when rank-1 is a false match, noise will be introduced to the new query, thus it's difficult to have a true match on rank-1 in the new ranking list. However, diversity will be introduced to the new query as well, and some true matches will have a higher rank in the new ranking list, so the performance of mAP has relatively higher improvement.

Two sample re-ID results on the Market-1501 dataset are shown in Fig. 3.5. It is clear that more true matches are found using BQE. Our method improves the baseline algorithm to a great extent.

Comparison with the state-of-the-art methods. On the Market-1501 dataset, we obtain rank-1 of 85.24%, mAP of 69.79% with the APR. We achieve the best rank-1 accuracy and mAP among the competing methods. On the DukeMTMC-reID dataset, we achieve rank-1 of 76.48% and mAP of 60.97% with APR and Euclidean distance when repeat BQE for 3 iterations. We achieve the best mAP among the competing methods, and the second best in rank-1 accuracy (the highest rank-1 accuracy is

Table 3.2 BQE results on DukeMTMC-reID. The best and second highest results are highlighted in blue and red.

Methods	rank-1	rank-5	rank-20	mAP
BoW+kissme [142]	25.13	-	-	12.17
LOMO+XQDA [54]	30.75	-	-	17.04
GAN (R, 702) [147]	67.68	-	-	47.13
SVDNet [96]	76.7	-	-	56.8
APR+EU	70.69	84.78	91.78	51.89
APR+EU+BQE	74.95	85.95	92.77	59.48
APR+EU+BQEI	76.48	85.81	91.49	60.97
APR+Kissme	70.78	84.15	91.15	50.62
APR+Kissme+BQE	75.49	85.456	91.60	55.79
APR+XQDA	71.18	84.02	91.24	51.21
APR+XQDA+BQE	74.77	85.32	91.60	58.43
IDE+EU	61.71	76.75	85.60	41.21
IDE+EU+BQE	70.37	81.73	89.22	53.03
IDE+Kissme	66.87	80.92	88.46	44.95
IDE+Kissme+BQE	69.12	80.25	88.06	48.60
IDE+XQDA	67.05	79.75	88.06	45.33
IDE+XQDA+BQE	70.33	81.28	88.82	51.72

reported by Sun *et al.* [96]). On the CUHK03 dataset, our method yields the best rank-1 accuracy result of 34.78% and the second best mAP of 34.46% (the highest mAP is reported by Zhong *et al.* [149]).

Effective of more iterations. We conduct BQE method for three iterations with the setting that produces the best performance. On these four datasets, BQE with more iterations (BQEI) shows superior results. On the Market-1501 dataset, BQEI achieves a better mAP of 70.74% (+0.95%) and a relatively lower rank-1 accuracy of 84.91% (-0.33%). On the DukeMTMC-reID dataset, compared with the BQE method, rank-1 increases from 74.95% to 76.48%, mAP increases from 59.48% to 60.97%. On the MARS dataset, an improvement of 1.98% on rank-1 accuracy and an improvement of 1.35% on mAP are observed. On the CUHK03 dataset, rank-1 accuracy and mAP of BQEI are 34.78% and 34.46%, respectively. An improvement of 0.93% on rank-1 and an improvement of 2.39% on mAP are observed.

3.4.3 Algorithm Analysis

Number of top ranked candidates. An important parameter to be considered is K, which defines how many candidates will be pooled to construct a new query together with the original query. As we can imagine, when K is 0, the new query is equal to the original query. When K becomes larger, more images in the ranking list

Table 3.3 BQE results on MARS. The best highest results are highlighted in blue.

Methods	rank-1	rank-5	rank-20	mAP
HistLBP+XQDA [141]	18.6	33.0	45.9	8.0
gBiCov+XQDA	9.2	19.8	33.5	3.7
LOMO+XQDA	30.7	46.6	60.9	16.4
BoW+Kissme	30.6	46.2	59.2	15.5
CNN+EU [141]	56.47	73.09	83.23	38.82
CNN+EU+BQE	59.32	77.18	87.19	42.75
CNN+Kissme [141]	63.54	79.35	88.4	45.20
CNN+Kissme+BQE	66.07	81.23	89.78	47.54
CNN+XQDA [141]	65.59	81.75	90.10	46.84
CNN+XQDA+BQE	66.46	82.25	90.39	51.38
CNN+XQDA+BQEI	68.42	81.59	88.82	52.73

Table 3.4 BQE results on CUHK03. The best results are highlighted in blue.

Methods	rank-1	rank-5	rank-20	mAP
k-reciprocal [149]	34.7	-	-	37.4
CNN+EU	21.35	37.50	57.42	19.75
CNN+EU+BQE	24.00	39.85	59.21	21.91
CNN+Kissme	27.88	46.57	65.71	26.80
CNN+Kissme+BQE	30.71	50.71	68.93	30.49
CNN+XQDA	29.50	50.00	71.35	28.14
CNN+XQDA+BQE	33.85	54.00	74.78	32.07
CNN+XQDA+BQEI	34.78	50.92	73.01	34.46

are used to generate the new query. To assign a suitable K value, we compare the result under different K values for different datasets. As for Market-1501 benchmark, the curve is shown in Fig. 3.6. At first, both recall and precision are improved. When K increases, mAP raises slowly and then remains the same value, rank-1 accuracy drops slowly and has a fluctuation. We set $K = 5$ to get a satisfying result.

Stability study. Fig. 3.7 presents the performance of three query expansion methods when have different K value. Notice that when K is small, the three algorithms produce similar results on both rank-1 accuracy and mAP. As K increases, the performance of BQE drops a little and remain a relative high accuracy. On the contrary, for AQE and MQE, both rank-1 accuracy and mAP drop rapidly.

We further investigate the weight of top 500 images in 3.8 . We observe that the weight of top ranked images drops rapidly, and remains nearly zero when $K > 50$. The Bayesian model plays an important role on control the impact of the top ranked images to form the expanded query.

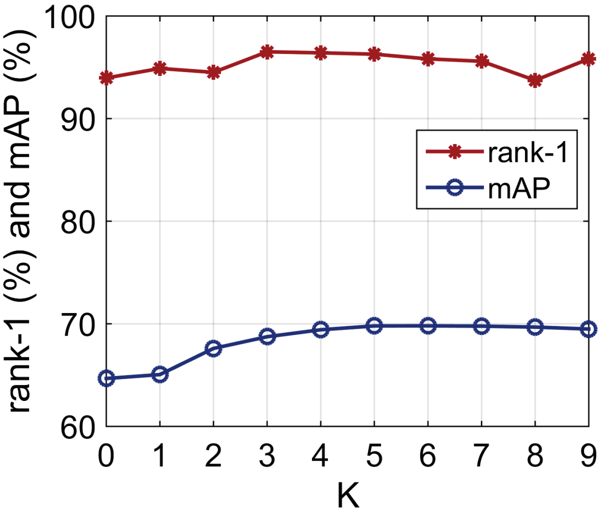


Fig. 3.6 mAP and CMC vs. *K* value changes on the Market-1501 dataset. The experiments are implemented using APR and Euclidean distance.

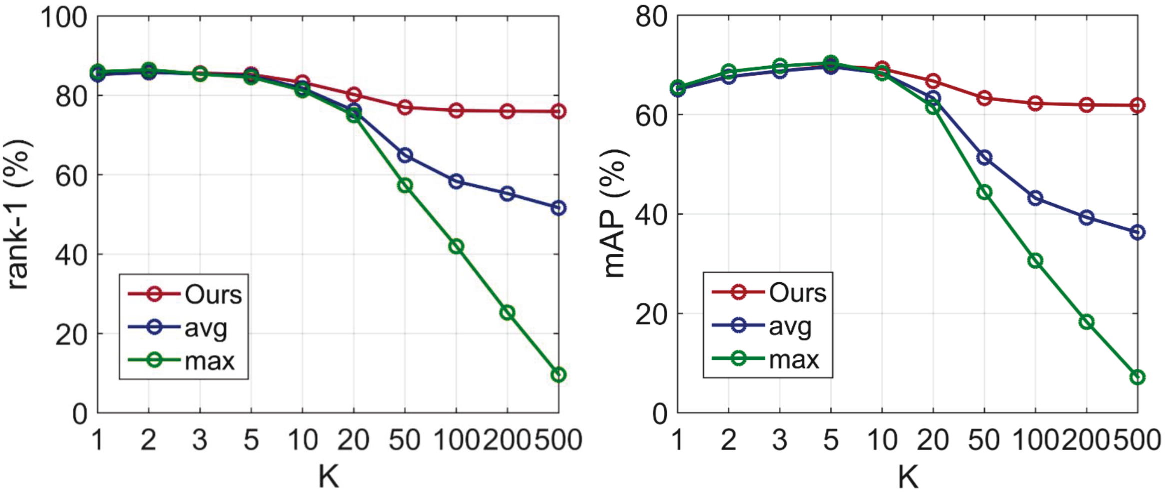


Fig. 3.7 mAP and CMC of three methods vs. *K* value changes on the Market-1501 dataset. The experiments are implemented using APR and Euclidean distance.

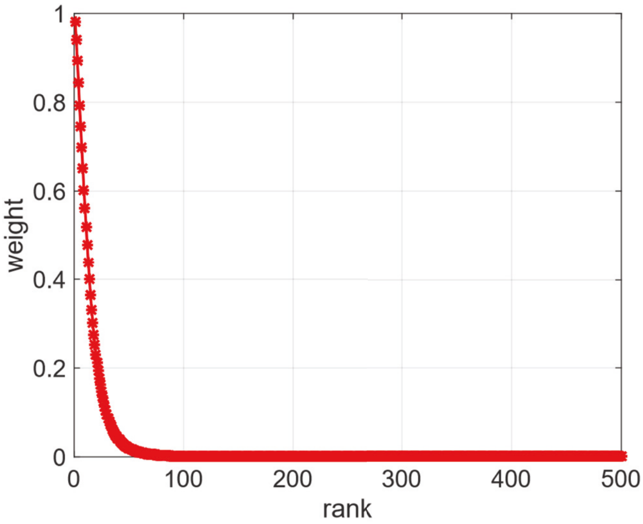


Fig. 3.8 Pooling weight of top 500 ranked images on Market-1501.

Number of iterations. The impact of the number of iteration T is shown in Fig. 3.9. On the one hand, when T increases, mAP increases slowly and then remains a relatively stable level. The reason is that as more images are merged into the query, the diversity of the query vector is enhanced, thus improving the retrieval recall.

On the other hand, although mAP is improved, more iterations of query expansion do not exactly lead to a higher rank-1 accuracy. From Fig. 3.9, on Market-1501, rank-1 accuracy obtains the best result in the first iteration, and then decreases to about 84.9%. For further iterations, rank-1 accuracy remains, which is still higher than the baseline. The main reason for this phenomenon is that some false matches might be ranked on the top, and are pooled into the new query several times during the iterations. That being said, on DukeMTMC-reID (Table 3.2), MARS (Table 3.3) and CUHK03 (Table 3.4), both rank-1 and accuracy improve with more iterations.

Overall speaking, mAP usually benefits from more iterations of BQE; rank-1 accuracy can be improved as well, but may be compromised in some cases due to the deteriorated query.

3.4.4 Comparison with Other Re-ranking Methods

In this section we implement and compare some query expansion methods such as AQE and MQE as we discussed in Section 3.3.3. Other kinds of re-ranking methods are also adopted and compared with BQE and the baseline.

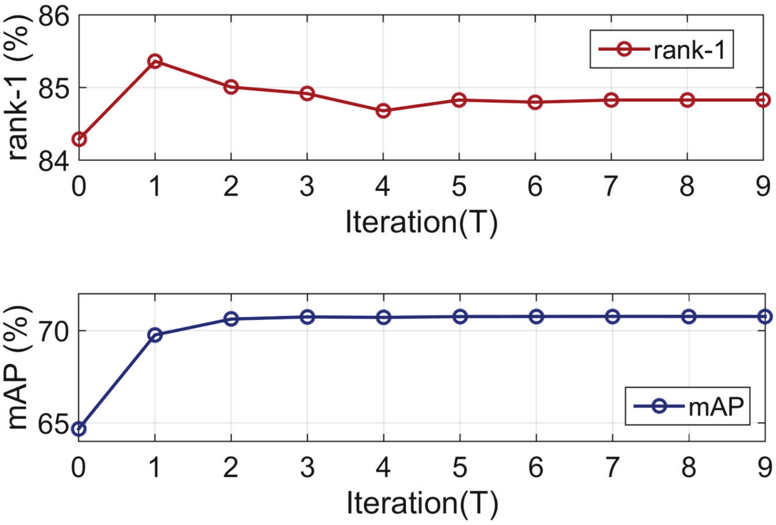


Fig. 3.9 Rank-1 and mAP vs. T value changes on the Market-1501 dataset. The experiments are implemented using APR and Euclidean distance.

- SVM query expansion. An SVM classifier is firstly trained on training data, then it replaces the Bayesian model in predicting labels of candidates in the initial ranking list.

- Graph based re-ranking method [135]. The top candidates' nearest neighborhoods are used as queries to get the ranking lists. Then a weighted undirected graph is built using the top candidates from those ranking lists. Here, the Jaccard similarity coefficient between two neighbors is used as weight.

- K-reciprocal Encoding [149]. This method calculates a k-reciprocal feature for re-ranking under the Jaccard distance. The final distance is computed by combining the original distance and the Jaccard distance. On the contrary, we propose a new query to compute the final distance. The time complexity of this method is $\mathcal{O}(NlogN)$, while ours is $\mathcal{O}(N)$. The two methods are complementary to each other. In Table 5 we show that we further improve the performance when build BQE upon the K-reciprocal Encoding [149].

The CMC and mAP results of different re-ranking methods are reported in Table 3.5. Here all the results are calculated using APR under Euclidean distance. We observe that all the re-ranking methods consistently improves the rank-1 accuracy and mAP over the original re-ID baseline. Among the query expansion methods (the second row to the six row in Table 3.5), the proposed BQE method achieves the best

Table 3.5 Re-ranking methods comparison on the Market-1501 dataset. Our method is more robust to parameters than average/max pooling. Our method also requires much less off-line computation and thus more flexible against database updates than [134, 149]. The best and second highest results are highlighted in blue and red.

Methods	rank-1	mAP
APR	84.29	64.67
APR+AQE	84.56	70.40
APR+MQE	84.88	69.60
APR+SVM	85.12	67.33
APR+Graph [134]	84.71	67.98
APR+k-reci [149]	85.87	77.27
APR+BQE	85.24	69.79
APR+BQEI	84.91	70.74
APR+Graph+BQE	85.42	70.35
APR+k-reci+BQE	86.57	76.95

result in rank-1 accuracy and BQE with three iterations achieves the best result in mAP. Other query expansion methods also beat the baseline to some extent.

The graph based method [134] exhibits a similar result to the query expansion methods. We also notice that the K-reciprocal encoding method achieves best result with 85.87% on rank-1 and 77.27% on mAP. However, the shortcoming of this method is that it needs much more offline training time, a process that needs to be re-computed every time the gallery is updated. We use BQE behind some other re-ranking methods, and the results are shown below. We also use BQE after these re-ranking method, and observe that when BQE is conducted after other graph based method, the rank-1 accuracy improves from 84.71% to 85.42%, and mAP improves from 67.89% to 70.30%. When we perform BQE after k-reciprocal method, rank-1 accuracy improves from 85.87% to 86.57.

3.5 Conclusion

We introduce a Bayesian query expansion algorithm to improve the recall of existing person Re-ID approaches. The pairwise similarity scores between images from the same and different identities are used to train the Bayesian model. For each query, the top ranked candidates in the initial ranking list are selected, and the features are pooled with the query using Bayesian model. A new query is then obtained and can be used to produce a new ranking list. The experiments show that our approach consistently improves the performance of baselines and is very robust to feature representation and metric learning methods. Our results are competitive with the state-of-the-art methods on four large-scale re-ID datasets.

In the future, we will further investigate effective and efficiency re-ranking methods. A promising direction is to integrate human labor in the re-ranking process [58, 104].

Chapter 4

Improving Person Re-identification by Attribute and Identity Learning

In this chapter, we exploit attribute information for person re-identification task to improve existing re-ID methods. Person re-identification (re-ID) and attribute recognition share a common target at the pedestrian description. Their difference consists in the granularity. Attribute recognition focuses on local aspects of a person while person re-ID usually extracts global representations. Considering their similarity and difference, in this chapter we propose a very simple convolutional neural network (CNN) that learns a re-ID embedding and predicts the pedestrian attributes simultaneously. This multi-task method integrates an ID classification loss and a number of attribute classification losses, and back-propagates the weighted sum of the individual losses.

Albeit simple, we demonstrate on two pedestrian benchmarks that by learning a more discriminative representation, our method significantly improves the re-ID baseline and is scalable on large galleries. We report competitive re-ID performance compared with the state-of-the-art methods on the two datasets.

4.1 Introduction

We propose the attribute-person recognition (APR) network to exploit both identity labels and attribute annotations for person re-ID. By combining the attribute recognition task and identity classification task, the APR network is capable of learning more discriminative feature representations for pedestrians, including global and local descriptions. Specifically, we take attribute predictions as additional cues for the identity classification. Considering the dependencies among pedestrian attributes, we first re-weight the attribute predictions and then build identification upon these re-weighted attributes descriptions. Comparing with previous re-ID literature which takes attributes into consideration, this method differs in two main aspects. First, in most previous re-ID methods, attributes are used to strengthen the relationship of image pairs, and their correlations are hardly considered. However, many attributes usually co-occur simultaneous on a person, and the correlations of attributes are helpful to re-weight the prediction of each attribute. For example, the attributes *skirt* and *handbag* are highly related to *female* rather than *male*. Given these gender-biased attribute descriptions, the probability of the attribute *female* should increase. We introduce an Attribute Re-weighting Module to utilize correlations among attributes and optimize the attribute predictions. Second, our work systematically investigates how person re-ID and attribute recognition benefit each other. On the one hand, identity labels provide global descriptions for person images, which have been proved effective to learn a good feature representation for a person in many re-ID works [7, 114, 142]. On the other hand, attribute labels provide detailed local descriptions. By exploiting both local (attribute) and global (identity) information, one is able to learn a better representation for a person, thereby achieving higher accuracy for person attribute recognition and person re-ID.

We evaluate the performance of the proposed method APR on two large-scale re-ID datasets. The experimental results show that our method achieves competitive re-ID accuracy to the state-of-the-art methods. In addition, we demonstrate that the proposed APR yields improvement in the attribute recognition task over the baseline in all the testing datasets. The main contributions are summarized as follows:

(1) We have manually labeled a set of pedestrian attributes for the Market-1501 dataset and the DukeMTMC-reID dataset. Attribute annotations of both datasets are publicly available on our website https://vana77.github.io.

(2) We propose a novel attribute-person recognition (APR) framework. It learns a discriminative Convolutional Neural Network (CNN) embedding for identities and attributes recognition.

(3) We introduce the Attribute Re-weighting Module, which corrects predictions of attributes based on the learned dependency and correlation among attributes.

(4) We achieve competitive accuracy in compared with the state-of-the-art re-ID methods on two large-scale datasets *i.e.*, Market-1501 [142] and DukeMTMC_reID [146]. We also demonstrate improvement in the attribute recognition task.

4.2 Related Work

Attributes for person re-ID. In some early attempts, attributes are used as auxiliary information to improve low-level features [44, 62, 91, 94]. In [43, 44], low-level descriptors and SVM are used to train attribute detectors, and the attributes are integrated by several metric learning methods. Su *et al.* [91, 94] utilize both low-level features and camera correlations learned from attributes for re-identification in a systematic manner. In [79], a dictionary learning model is proposed that exploits the discriminative attributes for the classification task. Recently, some deep learning methods are proposed. Franco *et al.* [19] propose a coarse-to-fine learning framework, which is comprised of a set of hybrid deep networks. The network is trained for distinguishing person/not person, predicting the gender of a person and person re-ID, respectively. In this work, the networks are trained separately and might overlook the complementarity of the ID label and the attribute label. Besides, gender is the only attribute used in the work, so that the correlation between attributes is not leveraged. However, these works do not consider the correlation between attributes nor show if the proposed method improves the attribute recognition baselines. In [93], Su *et al.* first train a network on an independent dataset with attribute label, and then fine-tune the network the target dataset using only identity label with triplet loss. Finally, the predicts attribute labels for the target dataset is combined with the independent dataset for the final round of fine-tuning. Similarly, in [88], the network is pre-trained on an independent dataset labeled with attributes, and then fine-tuned on another set with person ID. In [131], a set of attribute labels are used as the query to retrieve the person image. Adversarial learning is used to generate image-analogous concepts for query attributes and get it matched with the image in both the global level and semantic ID level. Wang *et al.* [105] propose an unsupervised re-ID method that shares the source domain knowledge through attributes learned from labelled source data and transfers such knowledge to unlabelled target data by a joint attribute identity transfer learning across domains.

4.3 Attribute Annotation

We manually annotate the Market-1501 [142] dataset and the DukeMTMC-reID [142] dataset with attribute labels. Although the Market-1501 and DukeMTMC-reID datasets are both collected in university campuses and most identities are students, they are significantly different in seasons (summer vs. winter) and thus have distinct clothes. For instance, many people wear dresses or pants in Market-1501 but most of the people

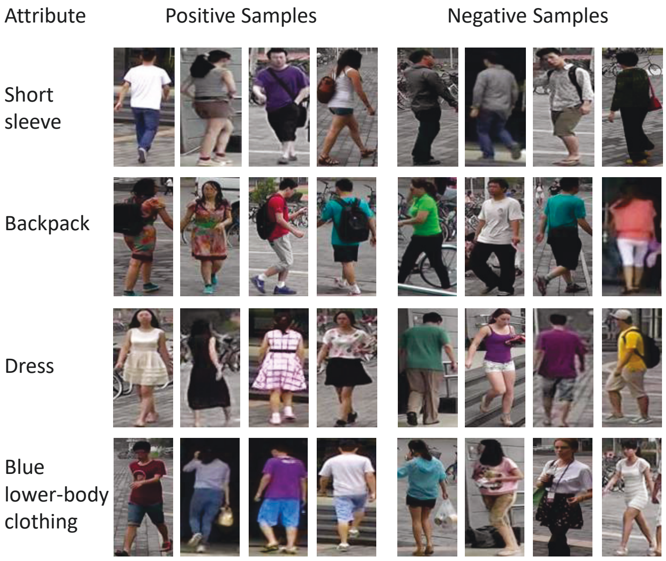
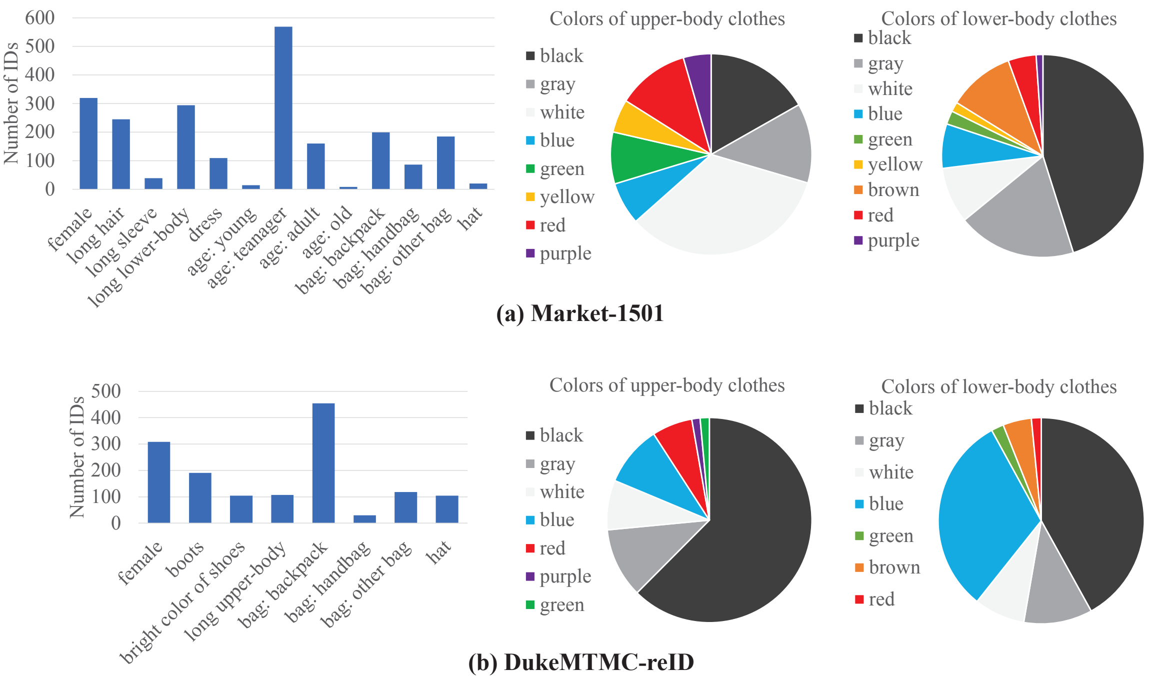


Fig. 4.1 Positive and negative examples of some representative attributes: *short sleeve*, *backpack*, *dress*, *blue lower-body clothing*.

wear pants in DukeMTMC-reID. So for the two datasets, we use two different sets of attributes. The attributes are carefully selected considering the characteristics of the datasets, so that the label distribution of an attribute (*e.g.*, wearing a hat or not) is not heavily biased.

For Market-1501, we have labeled 27 attributes: gender (male, female), hair length (long, short), sleeve length (long, short), length of lower-body clothing (long, short), type of lower-body clothing (pants, dress), wearing hat (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), carrying other types of bag (yes, no), 8 colors of upper-body clothing (black, white, red, purple, yellow, gray, blue, green), 9 colors of lower-body clothing (black, white, red, purple, yellow, gray, blue, green, brown) and age (child, teenager, adult, old). Positive and negative examples of some representative attributes of the Market-1501 dataset are shown in Fig. 4.1.

For DukeMTMC-reID, we have labeled 23 attributes: gender (male, female), shoe type (boots, other shoes), wearing hat (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), carrying other types of bag (yes, no), color of shoes (dark, bright), length of upper-body clothing (long, short), 8 colors of upper-body clothing (black, white, red, purple, gray, blue, green, brown) and 7 colors of lower-body clothing (black, white, red, gray, blue, green, brown).



(a) Market-1501

(b) DukeMTMC-reID

Fig. 4.2 The distributions of attributes on (a) Market-1501 and (b) DukeMTMC-reID. The left figure of each row shows the numbers of positive IDs for attributes except the color of upper/lower-body clothing. The middle and right pie chart illustrate the distribution of the colors of upper-body clothing and lower-body clothing, respectively.

Note that all the attributes are annotated at the identity level. For example, in Fig. 4.1, the first two images in the second row are of the same identity. Although we cannot see the backpack clearly in the second image, we still annotate there is a "backpack" in the image. For both Market-1501 and DukeMTMC-reID, we illustrate the attribute distribution in Fig 4.2. We define correlation of two attributes as the possibility that they co-occur on a person.

4.4 The Proposed Method

We first describe the necessary notations and two baseline methods in Section 4.4.1 and then introduce our proposed Attribute-Person Recognition network in Section 4.4.2.

4.4.1 Preliminaries

Let $\mathcal{S}_{\mathcal{I}} = \{(x_1, y_1), ..., (x_n, y_n)\}$ be the pedestrian identity labeled data set, where x_i and y_i denotes the i-th image and its identity label, respectively. For each image

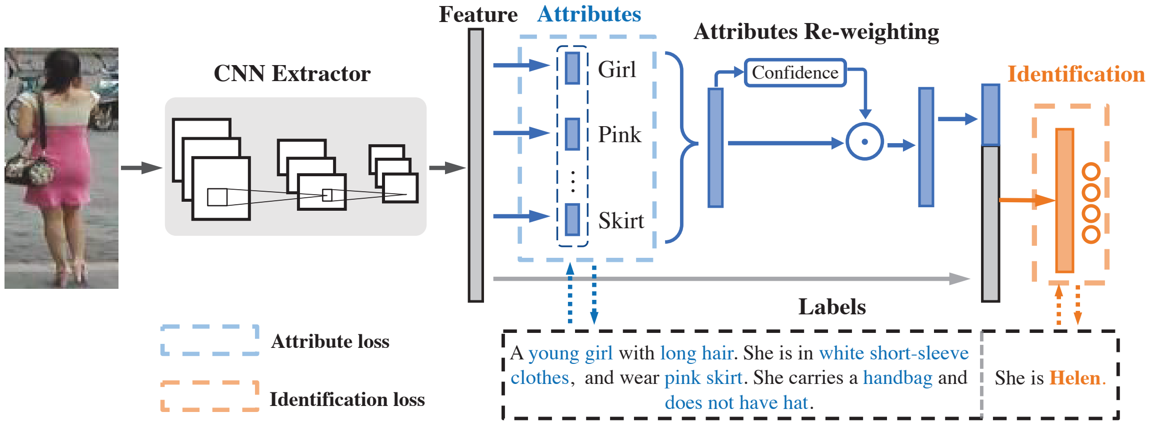


Fig. 4.3 An overview of the APR network. APR contains two classification part, one for attribute recognition and the other for identification. Given an input image, the person feature representation is extracted by the CNN extractor ϕ. Subsequently, the attribute classifiers predict attributes based on the image feature. Here we calculate the attribute classification losses by the attribute predictions and ground truth labels. For the identity classification part, we take the attribute predictions as additional cues. Specifically, we first re-weight the local attribute predictions by the Attribute Re-weighting Module and then concatenate them with the global image feature. The final identification is built upon the concatenated local-global feature.

$x_i \in \mathcal{S}_{\mathcal{I}}$, we have the attributes annotations $\boldsymbol{a}_i = (a_i^1, a_i^2, ..., a_i^m)$, where a_i^j is the j-th attribute label for the image x_i, and m is the number of attributes classes. Let $\mathcal{S}_{\mathcal{A}} = \{(x_1, \boldsymbol{a}_1), ..., (x_n, \boldsymbol{a}_n)\}$ be the attribute labeled set. Note that set $\mathcal{S}_{\mathcal{I}}$ and set $\mathcal{S}_{\mathcal{A}}$ share common pedestrian images $\{x_i\}$. Based on these two set $\mathcal{S}_{\mathcal{I}}$ and $\mathcal{S}_{\mathcal{A}}$, we have the following two baselines:

Baseline 1 ID-discriminative Embedding (IDE). Following recent works [114, 142, 146], we take IDE to train the re-ID model. IDE regards re-ID training process as an image identity classification task. It is trained only on the identity label data set $\mathcal{S}_{\mathcal{I}}$. We have the following objective function for IDE:

$$\min_{\boldsymbol{\theta}_I, \boldsymbol{w}_I} \sum_{i=1}^{n} \ell(f_I(\boldsymbol{w}_I; \phi(\boldsymbol{\theta}_I; x_i)), y_i), \tag{4.1}$$

where ϕ is the embedding function, parameterized by $\boldsymbol{\theta}_I$, to extract the feature from the data x_i. CNN models [142, 146] are usually used as the embedding function ϕ. f_I is an identity classifier, parameterized by $\boldsymbol{w}_I$, to classify the embedded image feature

$\phi(\boldsymbol{\theta}_I; x_i)$ into a k-dimension identity confidence estimation, in which k is the number of identities. ℓ denotes the suffered loss between classifier prediction and its ground truth label.

Baseline 2 Attribute Recognition Network (ARN). Similar to the IDE baseline for identity prediction, we propose the Attribute Recognition Network (ARN) for attribute prediction. ARN is trained only on the attribute label data set $\mathcal{S}_{\mathcal{A}}$. We define the following objective function for ARN:

$$\min_{\boldsymbol{\theta}, \boldsymbol{w}_A} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i)), a_i^j), \tag{4.2}$$

where f_{A_j} is the j-th attribute classifier, parameterized by $\boldsymbol{w}_{A_j}$, to classify the embedded image representation $\phi(\boldsymbol{\theta}; x_i)$ to the j-th attribute prediction. We take the sum of all the suffered losses for m attribute predictions on the input image x_i as the loss for the i-th sample.

In the evaluation stage of person re-ID task, for both baseline models, we use the embedding function $\phi(\boldsymbol{\theta}; \cdot)$ to embed the query and gallery images into the feature space. The query result is the ranking list of all gallery data according to the Euclidean Distance between the query data and each gallery data, *i.e.*., $||\phi(\boldsymbol{\theta}; x_q) - \phi(\boldsymbol{\theta}; x_g)||_2$, where x_q and x_g denote the query image and the gallery image, respectively. For the evaluation of attribute recognition task, we take the attribute prediction $f_A(\boldsymbol{w}_A; \phi(\boldsymbol{\theta}; \cdot))$ as the output, thereby evaluated with the ground truth by the classification metric.

4.4.2 Attribute-Person Recognition Network

Architecture Overview

The pipeline of the proposed APR network is shown in Fig. 4.3. APR network contains two prediction parts, one for attribute recognition task and the other for identity classification task. Given an input pedestrian image, the APR network first extracts the person feature representation by the CNN extractor ϕ. Subsequently, APR predicts attributes based on the image feature. Here we calculate the attribute losses by the attribute prediction and ground truth labels. For the identity classification part, motived by the fact that local descriptors (attributes) benefit global identification, we take the attribute predictions as additional cues for identity prediction. Specifically, we first re-weight the local attribute predictions by the Attribute Re-weighting Module and then concatenate them with the global image feature. The final identification is built upon the concatenated local-global feature.

Attribute Re-weighting Module

The Attribute Re-weighting Module is designed to utilize the dependencies and correlations among pedestrian attributes. The combination of attribute predictions contains inner connections among these local descriptions on a person, resulting in refining the prediction score of each attribute.

Suppose the attribute predictions for the image x_i are $\tilde{\boldsymbol{a}}_i = (\tilde{a}_i^1, \tilde{a}_i^2, ..., \tilde{a}_i^n)$, where $\tilde{a}_i^j$ is the j-th attribute prediction score from the attribute classifier f_{A_j}. For the j-th attribute, we learn the confidence score c_i^j for its prediction $\tilde{a}_i^j$ by the vertical concatenation of all the attributes prediction $\tilde{\boldsymbol{a}}_i$, *i.e.*,

$$c_i^j = \texttt{Sigmoid}(\boldsymbol{v}_j \tilde{\boldsymbol{a}}_i + b_j), \tag{4.3}$$

where $\boldsymbol{v}_j \in \mathbb{R}^{m\times 1}$ and b_j is the correlation vector and bias for the j-th attribute, respectively. The confidence score c_{i_j} represents the reliability of the prediction $\tilde{a}_i^j$. Therefore, we adjust the origin attribute prediction score to the re-weighted prediction by the confidence score, *i.e.*, $\hat{a}_i^j = c_i^j \times \tilde{a}_i^j$.

Optimization

To exploit the attributes data $\mathcal{S}_{\mathcal{A}}$ as auxiliary annotations for the re-ID task, we propose Attribute-Person Recognition (APR) network. The APR network is trained on the combined data set $\mathcal{S}$ of the identity set $\mathcal{S}_{\mathcal{I}}$ and the attribute set $\mathcal{S}_{\mathcal{A}}$, *i.e.*, $\mathcal{S} = \{(x_1, y_1, \boldsymbol{a}_1), ..., (x_n, y_n, \boldsymbol{a}_n)\}$. For a pedestrian image x_i, we first extract the image feature representation by the embedding function $\phi(\boldsymbol{\theta}; \cdot)$. Based on the image representation $\phi(\boldsymbol{\theta}; x_i)$, two object functions are optimized simultaneously:

The object function for attribute predictions. Similar to the baseline ARN, the attribute predictions are obtained by a set of attribute classifiers on the input image feature, *i.e.*, $\{f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i))\}$. We then optimize the object function for attribute predictions the same as Eqn. (4.2).

The object function for identification. To introduce the attributes into identity prediction, we gather the attribute predictions $\{f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i))\}$ and re-weight them by the Attribute Re-weighting Module. We combine the re-weighted attribute predictions $\hat{\boldsymbol{a}}_i$ and the image global feature $\phi(\boldsymbol{\theta}; x_i)$ to form a local-global representation.

The identity classification is built upon the new feature. Thus we have the following object function for identity prediction:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}_I} \sum_{i=1}^{n} \ell(\hat{f}_I(\hat{\boldsymbol{w}}_I; \hat{\boldsymbol{a}}_i, \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{4.4}$$

where $\hat{\boldsymbol{a}}_i = (\hat{a}_i^1, \hat{a}_i^2, ..., \hat{a}_i^j)$ is the concatenation of the re-weighted attribute predictions. $\hat{f}_I$ is the identity classier, parameterized by $\hat{\boldsymbol{w}}_I$, to predict the identity based on attribute predictions $\hat{\boldsymbol{a}}_i$ and image embeddings $\phi(\boldsymbol{\theta}; x_i)$.

The overall object function. Considering both attribute recognition and identity prediction, we define the overall object function as followings:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}_I,\boldsymbol{w}_A} \lambda \sum_{i=1}^{n} \ell(\hat{f}_I(\hat{\boldsymbol{w}}_I; \hat{\boldsymbol{a}}_i, \phi(\boldsymbol{\theta}; x_i)), y_i) + \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i)), a_i^j), \tag{4.5}$$

where λ is a hyper-parameter to balance the identity classification loss and the attribute recognition losses. We empirically discuss the effectiveness of λ in Section.4.5.1.

4.4.3 Implementation Details

We take ResNet-50 [27] with the last classification layer removed as the feature extractor $\phi(\theta; \cdot)$ for most experiments. The feature vector after the pool5 layer in ResNet-50 is taken as the output of extractor ϕ. For some experiments based on the CaffeNet [114], similarly, we take the feature vector after the fc7 layer as the output of extractor ϕ. We initialize the embedding extractor by ImageNet [86] pre-trained models. For the identity classifier $f_I(\boldsymbol{w}_I; \cdot)$, we take as input the embedded feature from $\phi(\theta; \cdot)$. The feature is then processed by a 512-dim fully-connected (FC) layer with Batch Normalization. We add a dropout layer with drop rate 0.5 for regularization. Finally, the classification layer with k class nodes is used to predict the identity. For the j-th attribute classifier $f_{A_j}(\boldsymbol{w}_{A_j}; \cdot)$, we simply adopt a FC layer for each attribute prediction. When evaluating the APR network for re-ID task, we take the vertical concatenation of the mbedded feature $\phi(\theta; x_i)$ and attribute predictions $\{f_{A_j}(\boldsymbol{w}_{A_j}; x_i)\}$ as the final feature representation for image x_i.

Following [146], we adopt the similar training strategy. Specifically, when using ResNet-50, we set the number of epochs to 60. The batch size is set to 32. Learning rate is initialized to 0.001 and changed to 0.0001 in the last 20 epochs. For CaffeNet, the number of epochs is set to 110. For the first 100 epochs, the learning rate is 0.1

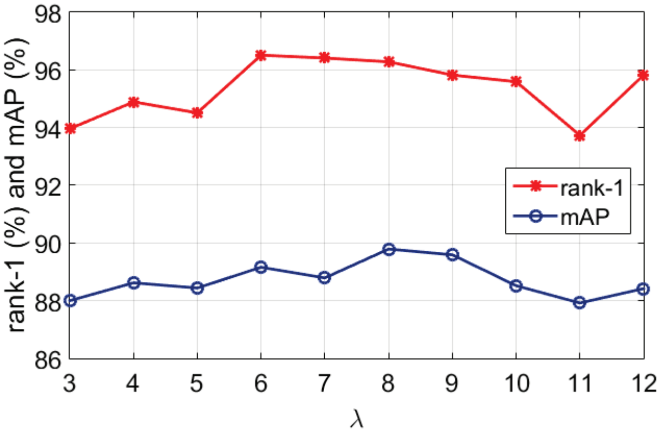


Fig. 4.4 The re-ID performance (rank-1 accuracy and mAP) curves on the validation set of Market-1501 with different values of the parameter λ in Eqn. (4.5). According to the performance curves, we set $\lambda = 8$ for all the other experiments on Market-1501, DukeMTMC-reID and PETA.

and changed to 0.01 in the last 10 epochs. The batch size is set to 128. Randomly cropping and horizontal flipping are implied on the input images during training.

4.5 Experimental Results

4.5.1 Evaluation of Person Re-ID task

Parameter validation. We first validate the parameter λ of APR on the validation set of Market-1501. λ is a key parameter balancing the contribution of identification loss and attribute recognition loss (Eq. 4.5). When $\lambda = 0$, the APR network reduces to Baseline 2 (ARN). When λ becomes larger, person identity classification will play a more important role. Re-ID performance result on the validation set of Market-1501 with different values of the parameter λ is presented in Fig. 4.4. As λ increases, we observe that both the rank-1 accuracy and mAP of the model first increase and then decrease. The best re-ID performance is obtained when $\lambda = 8$. Therefore, we use $\lambda = 8$ for APR in all the following experiments.

Attribute recognition improves re-ID over the baselines. Results on the three datasets are shown in Table 4.1 Table 4.2 and Table 4.3.

First, we observe that Baseline 2 (ARN) yields decent re-ID performance, *e.g.*, a rank-1 accuracy of 49.76% using ResNet-50 on Market-1501. Note that Baseline 2 only

Table 4.1 Comparison with state of the art on Market-1501. "C" and "R" represent CaffeNet and ResNet-50, *resp.* "w/o ARM" denote APR without the attribute re-weighting module. The numbers in the bracket are the number of training IDs.

Methods	Publish	rank-1	rank-5	rank-10	mAP
MBC [99]	(AVSS2017)	45.56	67	76	26.11
SML [36]	(ECCV2016)	45.16	68.12	76	-
SL [6]	(CVPR2016)	51.9	-	-	26.35
Attri [75]	(ICPR2016)	58.84	-	-	33.04
S-CNN [100]	(ECCV2016)	65.88	-	-	39.55
GAN [147]	(ICCV2017)	79.33	-	-	55.95
2Stream [146]	(TOMM2017)	79.51	90.91	94.09	59.87
Cont-aware [48]	(CVPR2017)	80.31	-	-	57.53
Part-align [137]	(ICCV2017)	81.0	92.0	94.7	63.4
SVDNet [96]	(ICCV2017)	82.3	92.3	95.2	62.1
Baseline 1 (C)	-	54.76	73.28	82.04	28.75
Baseline 1 (R)	-	80.16	92.03	94.98	57.82
Baseline 2 (R)	-	49.76	70.07	77.767	23.95
APR (C)	-	59.32	78.26	85.03	32.85
APR (R, w/o ARM)	-	85.71	94.32	96.46	66.59
APR (R)	-	**87.26**	**95.03**	**96.82**	**66.94**

utilizes attribute annotations without ID labels. This illustrates that attributes are capable of discriminating between different persons.

Second, by integrating the advantages in Baseline 1 and Baseline 2, our method exceeds the two baselines by a large margin. For example, when using ResNet-50, the rank-1 improvement on Market-1501 over Baseline 1 and Baseline 2 is 7.1% and 38.5%, respectively. On DukeMTMC-reID, APR achieves 9.7% and 27.78% improvement over Baseline 1 and Baseline 2 in rank-1 accuracy. Consistent finding also holds for PETA, *i.e.*, we observe improvements of 4.15% and 14.65% over Baseline 1 and Baseline 2 in rank-1 accuracy, respectively. This demonstrates the complementary nature of the two baselines, *i.e.*, identity and attribute learning.

Third, for both backbone models (*i.e.,* CaffeNet and ResNet-50), APR yields consistent improvement. On Market-1501, we obtain 4.56% and 7.1% improvements in rank-1 accuracy over Baseline 1 with CaffeNet and ResNet-50, respectively.

The effectiveness of the Attribute Re-weighting Module. We test APR with and without Attribute Re-weighting Module on the three re-ID datasets, and the results are shown in Table 4.1, Table 4.2 and Table 4.3. We observe performance improvement by using the Attribute Re-weighting Module for all the datasets. For Market-1501 with ResNet-50 as the backbone, the rank-1 and mAP improvements are 1.55% and 0.35%, respectively. For DukeMTMC-reID, the improvements are 0.36%

Table 4.2 Comparison with the state of the art on DukeMTMC-reID with ResNet-50. Rank-1 accuracy (%) and mAP (%) are shown. "w/o ARM" denotes APR without the Attribute Re-weighting Module.

Methods	rank-1	mAP
BoW+kissme [142]	25.13	12.17
LOMO+XQDA [54]	30.75	17.04
AttrCombine [75]	53.87	33.35
GAN [147]	67.68	47.13
SVDNet [96]	**76.7**	**56.8**
Baseline 1	64.22	43.50
Baseline 2	46.14	24.17
APR (w/o ARM)	73.56	54.79
APR	73.92	55.56

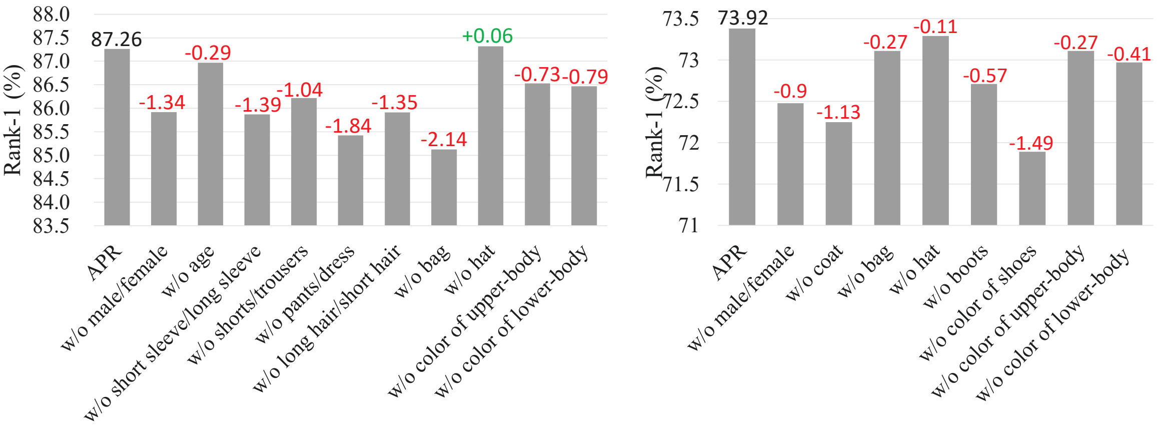
Table 4.3 Person reID performance on PETA with ResNet-50. Rank-1 accuracy (%) and mAP (%) are shown. "w/o ARM" denotes APR without the Attribute Re-weighting Module.

Methods	rank-1	rank-5	rank-10	rank-20	mAP
Baseline 1	53.90	68.32	73.04	80.14	49.60
Baseline 2	43.30	60.08	70.23	77.40	39.52
APR (w/o ARM)	56.91	72.10	78.48	83.45	53.81
APR	**58.05**	**78.34**	**75.73**	**84.01**	**55.84**

and 0.74%, respectively. For PETA, we observe improvements of 1.14% and 2.03% in rank-1 and mAP, respectively. The improvement is consistent on all experiments.

Comparison with the state-of-the-art methods. The comparison with the state-of-the-art algorithms on Market-1501 and DukeMTMC-reID is shown in Table 4.1 and Table 4.2, respectively. On Market-1501, we obtain **rank-1 = 87.26%, mAP = 66.94%** by APR using the ResNet-50 model. We achieve the best rank-1 accuracy and mAP among the competing methods. On DukeMTMC-reID, our results are **rank-1 = 73.92% and mAP = 55.56%** by APR using ResNet-50. Our method is thus shown to compare favorably with the state-of-the-art methods.

Ablation study of attributes. We evaluate the contribution of individual attributes on the re-ID performance. We remove each attribute from the APR system at one time, and the results on the two datasets are summarized in Fig. 4.5. We find that most of the attributes on Market-1501 and DukeMTMC-reID are indispensable. The most influencing attributes on the two datasets are *bag types* and *the color of shoes*,



(a) Market-1501

(b) DukeMTMC-reID

Fig. 4.5 Re-ID rank-1 accuracy on Market-1501 and DukeMTMC-reID. We remove one attribute from the system at a time. All the colors of upper-body clothing are viewed as one attribute here; the same goes for colors of lower-body clothing. Accuracy changes are indicated above the bars.

which lead to a rank-1 decrease of 2.14% and 1.49% on the two datasets, respectively. This indicates that pedestrians of the two datasets have different appearances. The attribute of "wearing a hat or not" seems to exert a negative impact on the overall re-ID accuracy, but the impact is very small.

Accelerating the retrieval process. Attributes can be used to speed up the evaluation process by filtering the gallery data based on attribute predictions. The gallery features and attribute predictions are usually calculated in advance (off-line). We take those attributes with higher prediction scores over a threshold as the reliable attributes. When an on-line query starts, we only take the gallery data with the same reliable attributes into consideration. Gallery images with different attribute predictions are filtered. Fig. 4.6 illustrates the re-ID performance over different percentages of remaining gallery data. As the percentage of remaining gallery data decreases from 42% to 9.7%, *i.e.*, accelerating retrieval from 2.4 times to 10.3 times, the rank-1 accuracies for re-ID decrease very gently. When we try a more aggressive speedup, the performance drops quickly. For example, we observe an accuracy drop of 22.78% when we speed up 193 times. Note that with remaining 9.74% gallery data, we still achieve 86.03% on rank-1 accuracy, which is very close to the original result 87.26%. It shows that APR could speed up the retrieval process 10.3 times with only a slight accuracy drop of 1.23%.

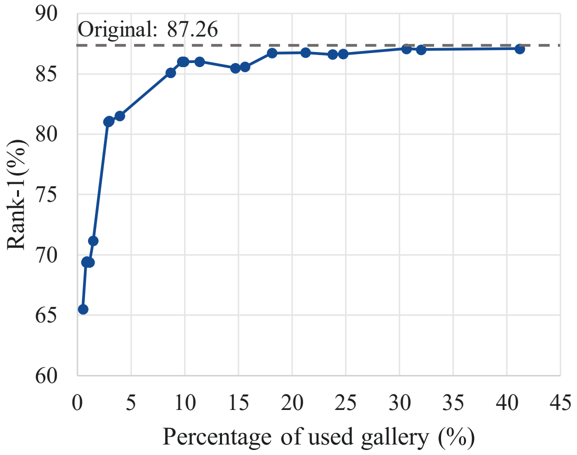


Fig. 4.6 Re-ID rank-1 accuracy curve on Market-1501 when using attribute to accelerate the retrieval process. For an on-line query, we only take the gallery data with the same reliable attributes into consideration. X-axis stands for different percentages of the *remaining* gallery data when using different filter threshold values. Note that APR could speed up the retrieval process 10.3 times (remaining 9.74% gallery data) with only a slight accuracy drop of 1.23%.

Table 4.4 Attribute recognition accuracy on Market-1501. In “APR”, parameter λ is optimized in Fig. 4.4. “L.slv”, “L.low”, “S.clth”, “B.pack”, “H.bag”, “C.up”, “C.low” denote *length of sleeve*, *length of lower-body clothing*, *style of clothing*, *backpack*, *handbag*, *color of upper-body clothing* and *color of lower-body clothing*, *resp.* “B2” denotes Baseline 2 (ARN).

	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	Avg
B2	87.5	85.8	84.2	93.5	93.6	93.6	86.6	88.1	78.6	97.0	72.4	71.7	86.0
APR	88.9	88.6	84.4	93.6	93.7	92.8	84.9	90.4	76.4	97.1	74.0	73.8	86.6

4.5.2 Evaluation of Attribute Recognition

We test attribute recognition on the galleries of the Market-1501, DukeMTMC-reID, PETA in Table 4.4, Table 4.5, and Table 4.6, respectively. We also evaluate our method on the CUB_200_2011 dataset, which contains 11,788 images of 200 bird classes. Each category is annotated with 312 attributes, which are divided into 28 groups and are used as 28 multi-class attributes in our experiments. The result is shown in Table 4.7. By comparing the results of APR and Baseline 2 (ARN), two conclusion can be drawn:

First, on all datasets, the overall attribute recognition accuracy is improved by the proposed APR network to some extent. The improvement is 0.26%, 0.08%, 0.2% and

Table 4.5 Attribute recognition accuracy on DukeMTMC-reID. “L.up”, “B.pack”, “H.bag”, “C.shoes”, “C.up”, “C.low” denote *length of sleeve*, *backpack*, *handbag*, *color of shoes*, *color of upper-body clothing* and *color of lower-body clothing*, *resp.* “B2” denotes Baseline 2 (ARN).

	gender	hat	boots	L.up	B.pack	H.bag	bag	C.shoes	C.up	C.low	Avg
B2	82.0	85.5	88.3	86.2	77.5	92.3	82.2	87.6	73.4	68.3	82.3
APR	84.2	87.6	87.5	88.4	75.8	93.4	82.9	89.7	74.2	69.9	83.4

Table 4.6 Attribute recognition accuracy on PETA.

Attribute	mean accuracy
MRFr2 [12]	71.1
ACN [95]	81.15
MVA [88]	84.61
Baseline 2	84.45
APR	**84.94**

Table 4.7 Attribute recognition accuracy on CUB_200_2011.

Methods	mean accuracy
Baseline 2	87.31
APR	**89.12**

1.58% on Market-1501, DukeMTMC-reID, PETA and CUB_200_2011, respectively. So overall speaking, the integration of identity classification introduces some degree of complementary information and helps in learning a more discriminative attribute model. Also, note that we achieve the best attribute recognition result on PETA among the state-of-the-art.

Second, we observe that the recognition rate of some attributes decreases for APR, such as *hair* and *B.pack* in Market-1501. However, Fig. 4.5 demonstrates that these attributes are necessary for improving re-ID performance. The reason probably lies in the multi-task nature of APR. Since the model is optimized for re-ID (Fig. 4.4), ambiguous images of certain attributes may be incorrectly predicted. Nevertheless, the improvement on the two datasets is still encouraging and further investigations should be critical.

4.6 Conclusions

In this chapter, we mainly discuss how re-ID is improved by the integration of attribute learning. Based on the complementarity of attribute labels and ID labels, we propose an attribute-person recognition (APR) network, which learns a re-ID embedding and predicts the pedestrian attributes under the same framework. We systematically investigate how the person re-ID and attribute recognition benefit each other. In addition, we re-weight the attribute predictions considering the dependencies and correlations among attributes of a person. We propose the attribute-person recognition (APR) network which learns a discriminative embedding for person re-ID and can make attribute predictions. The APR network contains both the ID classification and attribute classification losses which are respectively contained in the re-ID and attribute recognition baselines. To show the effectiveness of our method, we have annotated attribute labels on two large-scale re-ID datasets. The experimental results on two large-scale re-ID benchmarks demonstrate that by learning a more discriminative representation, APR achieves competitive re-ID performance compared with the state-of-the-art methods. We additionally use APR to accelerate the retrieval process of re-ID ten times with a minor accuracy drop of 1.26% on Market-1501. For attribute recognition, we also observe an overall precision improvement using APR.

Chapter 5

One-Example Video-Based Person Re-Identification by step-wise Learning

In this chapter, we focus on the one-example learning for video-based person re-Identification (re-ID). Unlabeled tracklets for the person re-ID tasks can be easily obtained by pre-processing, such as pedestrian detection and tracking. In this chapter, we propose an approach to exploiting unlabeled tracklets by gradually but steadily improving the discriminative capability of the Convolutional Neural Network (CNN) feature representation via step-wise learning. We first initialize a CNN model using one labeled tracklet for each identity. Then we update the CNN model by the following two steps iteratively: 1. sample a few candidates with most reliable pseudo labels from unlabeled tracklets; 2. update the CNN model according to the selected data. Instead of the static sampling strategy applied in existing works, we propose a progressive sampling method to increase the number of the selected pseudo-labeled candidates step by step. We systematically investigate the way how we should select pseudo-labeled tracklets into the training set to make the best use of them. Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 21.46 points (absolute, *i.e.*, 62.67% vs. 41.21%) on the MARS dataset, and 16.53 points on the DukeMTMC-VideoReID dataset.

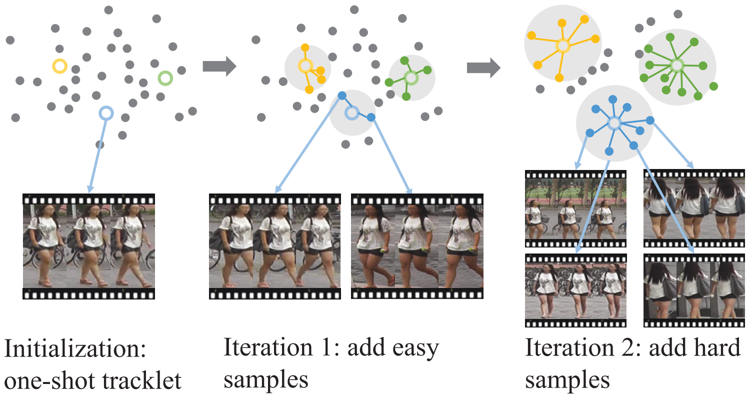


Fig. 5.1 An illustration of the unlabeled data sampling procedure in the feature space. The hollow point and solid point denote the labeled tracklet and unlabeled tracklet, respectively. The pseudo label of each unlabeled tracklet is assigned by its nearest labeled neighbor (indicated by the colored line). Different colors represent different identities. Samples in the shade will be incorporated into training. We adopt the easy and reliable pseudo-labeled tracklets for updating at the beginning and difficult ones in subsequence.

5.1 Introduction

Person re-identification (re-ID) aims at spotting the person-of-interest from different cameras. In recent years, person re-ID on the large-scale video data, such as surveillance videos, has attracted significant attention [29, 65, 106, 121, 136]. Most proposed approaches rely on the fully annotated data, *i.e.*., the identity labels of all the tracklets from multiple cross-view cameras. However, it is impractical to annotate very large-scale surveillance videos due to the dramatically increasing cost. Therefore, semi-supervised methods [66, 129] are of particular interest. This work mainly focuses on the one-example setting, in which only one tracklet is labeled for each identity.

The key challenge for the one-example video-based person re-ID is the label estimation for the abundant unlabeled tracklets [16, 129]. A typical approach is to generate the pseudo labels for the unlabeled data at first. The initial labeled data and some selected pseudo-labeled data are considered as an enlarged training set. Lastly, this new training set is adopted to train the re-ID model.

Most existing methods employ a static strategy to determine the quantity of selected pseudo-labeled data for further training. For example, Fan *et al.*. [16] and Ye *et al.*. [129] compare the prediction confidences of pseudo-labeled samples with a *pre-defined* threshold. The samples with higher confidence over the fixed threshold are then selected for the subsequent training. During iterations, these algorithms select a fixed and large number of pseudo-labeled data from beginning to end. However, it is inappropriate to keep the threshold fixed in the one-example setting. In this case, the initial model may be not robust due to the very few training samples. Only a few of pseudo-label predictions are reliable and accurate at the initial stage. If one still selects the same number of data as that in the later stages, it will inevitably involve many unreliable predictions. Updating the model with excessive not-yet-reliable data would hinder the subsequent improvement of the model.

In this chapter, to better exploit the unlabeled data in one-example video-based person re-ID, we propose the step-wise learning method **EUG** (**E**xploit the **U**nknown **G**radually). Initially, a CNN model is trained on the one-example labeled tracklet. EUG then iteratively updates the CNN by two steps, the label estimation step and the model update step. In the first step, EUG generates the pseudo labels for unlabeled tracklets, and selects some of pseudo-labeled tracklets for training according to the prediction reliability. The selected subset is continuously enlarged during iterations according to a sampling strategy. In the second step, EUG re-trains the CNN model on both the labeled data and the sampled pseudo-labeled subset. Particularly, as illustrated in Figure 5.1, EUG starts with a small-size subset of pseudo-labeled tracklets, which includes only the most reliable and easiest ones. In the subsequent stages, it gradually selects a growing number of pseudo-labeled tracklets to incorporate more difficult and diverse data. This is different from existing methods [66, 129].

To characterize the proposed progressive approach in one-example person re-ID, we intensively investigate two significant aspects, *i.e.*, how the progressive sampling strategy benefits the label estimation and which sampling criterion is effective for the confidence estimation in person re-ID. For the first aspect, we find that if we enlarge the sampled subset of pseudo-labeled data in a more conservative way (at a slower speed), the model achieves a better performance. If we enlarge the subset in a more aggressive way (at a faster speed), the model achieves a worse performance. Note that the previous static sampling strategy can be viewed as an extremely aggressive manner. For the second aspect, we investigate the gap between the classification measures and retrieval evaluation metrics. We find that the sampling criteria highly affect the performance of the proposed method. Instead of the classification measures, a distance-based sampling criterion for the reliability estimation may yield promising performance in person re-ID.

Our contributions are summarized as follows:

- We propose a progressive method for one-example video-based person re-ID to better exploit the unlabeled tracklets. This method adopts a dynamic sampling strategy to uncover the unlabeled data. We start with reliable samples and gradually include diverse ones, which significantly makes the model robust.
- We apply a distance-based sampling criterion for label estimation and candidates selection to remarkably improve the performance of label estimation.
- Our method achieves surprisingly superior performance on the one-example setting, outperforming the state-of-the-art by 21.46 points (absolute) on MARS and 16.53 points (absolute) on DukeMTMC-VideoReID.

5.2 Related Works

Extensive works have been reported to address the video-based person re-ID problem. One simple solution is using image-based re-ID methods, and obtaining video representations by pooling the frame features [29, 48, 65].

Supervised Video-based Person Re-ID. Recently, a number of deep learning methods are developed [76, 110, 121, 136, 150, 152]. The typical architecture is to combine CNN and RNN to learn a video representation or the similarity score. In [152], temporal attention information and spatial recurrent information are used to explore contextual representation. Another commonly used architecture is the Siamese network architecture [60, 110, 121], which also achieve reasonably good performance.

Semi-Supervised Video-based Person Re-ID. Most works of semi-supervised person re-ID are based on image [3, 18, 61, 69]. The approaches of these works include dictionary learning, graph matching, metric learning, *etc.* To the best of our knowledge, there are three works aiming at solving the semi-supervised video-based re-ID task. Zhu *et al.*. [155] proposed a semi-supervised cross-view projection-based dictionary learning (SCPDL) approach. A limitation is that this approach is only suitable for datasets that only captured by two cameras.

There are two recent works designed for one-example video re-ID task [66, 129]. Although [66, 129] claim them as *unsupervised* methods, they are *one-example* methods in experiments, as they require at least one labeled tracklet for each identity. They assume that the tracklets are obtained by tracking, and this process is automatic and unsupervised. Different tracklets from one camera with a long-time interval are assumed representing different identities. However, to conduct experiments in existing datasets, both methods require the annotation of at least a sample for each identity. To be more rigorous, we take this problem as a one-example task. Ye *et al.*. [129] propose a dynamic graph matching (DGM) method, which iteratively updates the image graph and the label estimation to learn a better feature space with intermediate estimated labels. Liu *et al.*. [66] update the classifier with K-reciprocal Nearest Neighbors (KNN) in the gallery set, and refine the nearest neighbors by apply negative sample mining with KNN in the query set. While graph-based semi-supervised learning [126] could possibly be adopted for one-example person Re-ID, it is time-consuming to solve a linear system for each query.

Progressive Paradigm. Curriculum Learning (CL) is proposed in [5], which progressively obtains knowledge from easy to hard samples in a pre-defined scheme. Kumar *et al.*. [42] propose Self-Paced Learning (SPL) which takes curriculum learning as a regularization term to update the model automatically. The self-paced paradigm is theoretically analyzed in [35, 70]. Some works manage to apply the progressive paradigm in the computer vision area [14]. We are inspired by these progressive algorithms. Compared with the existing SPL and CL algorithms, we incorporated the retrieval measures (the distance in feature space) into the learning mechanism, which well fits the evaluation metric for person re-ID. Moreover, most previous SPL and CL works mainly focus on the supervised and semi-supervised task. Few are used in the one-example learning setting.

5.3 The Progressive Model

5.3.1 Preliminaries

We first introduce the necessary notations. Let $\mathcal{L} = \{(x_1, y_1), ..., (x_{n_l}, y_{n_l})\}$ be the labeled dataset, and $\mathcal{U} = \{(x_{n_l+1}), ..., (x_{n_l+n_u})\}$ be the unlabeled dataset, where x_i and y_i denotes the i-th tracklet data and its identity label, respectively. We thus have $|\mathcal{L}| = n_l$ and $|\mathcal{U}| = n_u$ where $|\cdot|$ is the cardinality of a set. Following recent works [16, 56, 149], we take the training process as an identity classification task. For training on the labeled dataset, we have the following objective function:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=1}^{n_l} \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{5.1}$$

where ϕ is an embedding function, parameterized by $\boldsymbol{\theta}$, to extract the feature from the data x_i. CNN models [27, 31, 116, 118] are usually used as the function ϕ. f

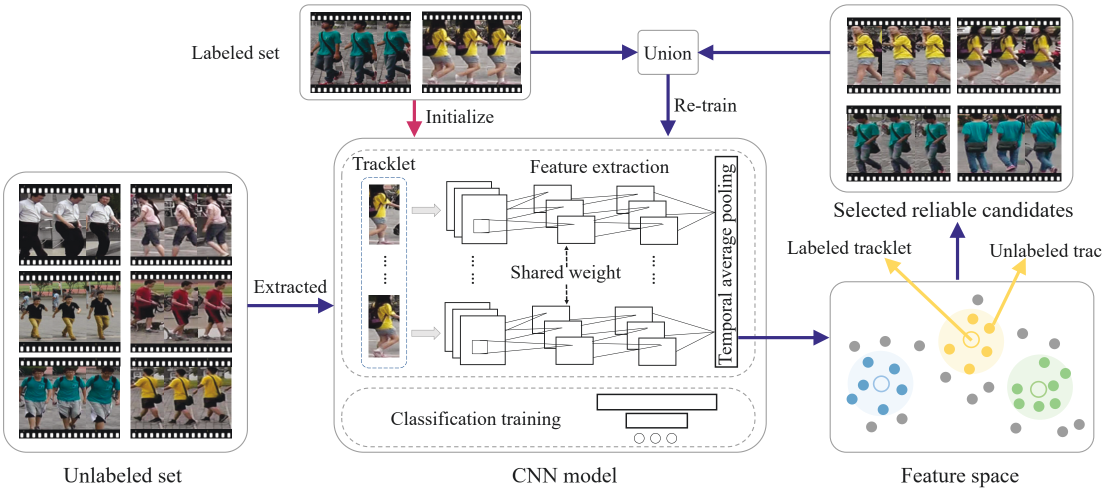


Fig. 5.2 Overview of the framework. Different colors represent different identity samples. The CNN model is initially trained on the labeled one-example data. For each iteration, we (1) select the unlabeled samples with reliable pseudo labels according to the distance in feature space and (2) update the CNN model by the labeled data and the selected candidates. We gradually enlarge the candidates set to incorporating more difficult and diverse tracklets. For a tracklet, each frame feature is first extracted by the CNN model and then temporally averaged as the tracklet feature. We take the training process as an identity classification task, and regard the evaluation as a retrieval problem on the features of the test tracklets.

is a function, parameterized by $\boldsymbol{w}$, to classifier the embedded feature $\phi(\boldsymbol{\theta};x_i)$ into a k-dimension confidence estimation, in which k is the number of identities. ℓ denotes the suffered loss on the label prediction $f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i))$ and its ground truth identity label y_i.

To exploit abundant unlabeled tracklets with pseudo labels, we consider the following objective function in the one-example re-ID problem:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w},s_i,\hat{y}_i} &\sum_{i=1}^{n_l} \ell(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),y_i)+\\ &\sum_{i=n_l+1}^{n_l+n_u} s_i\ell(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),\hat{y}_i)+R(\boldsymbol{s}),\end{aligned} \tag{5.2}$$

where $\hat{y}_i$ denotes the machine generated pseudo labels for the i-th unlabeled data. $s_i \in \{0,1\}$ is the selection indicator for the unlabeled sample x_i, which determine whether the suffered loss of pseudo-labeled data $(x_i,\hat{y}_i)$ is adopted in optimizing. We use $\boldsymbol{s}$ to indicate the vertical concatenation of all s_i. $R(\boldsymbol{s})$ is a regularization term on $\boldsymbol{s}$ to encourage incorporating more pseudo-labeled data.

In the evaluation stage, for both of query data and gallery data, we only use $\phi(\boldsymbol{\theta};\cdot)$ to embed each tracklet into the feature space. The query result is the ranking list of all gallery data according to the Euclidean Distance between the query data and each gallery data, *i.e..*, $||\phi(\boldsymbol{\theta};x_q)-\phi(\boldsymbol{\theta};x_g)||_2$, where x_q and x_g denote the query tracklet and the gallery tracklet, respectively.

5.3.2 Framework Overview

In this work, we propose a step-wise learning method to exploit the unlabeled data gradually and steadily. We adopt an alternative algorithm to solve the Eq. (6.5). Specifically, we first optimize $\boldsymbol{\theta}$ and $\boldsymbol{w}$, and then optimize $\hat{y}$ and $\boldsymbol{s}$, *i.e..*, the model updating and the label estimating.

Let $\mathcal{S}$ denote the set of selected pseudo-labeled candidates. We can obtain $\mathcal{S}$ by:

$$\mathcal{S} = \{(x_i,\hat{y}_i)|s_i = 1, n_l + 1 \leq i \leq n_l + n_u\}. \tag{5.3}$$

Our approach first trains an initial model on the labeled data $\mathcal{L}$, and then the initial model is applied to predict pseudo labels $\hat{y}$ on the unlabeled data. In subsequence, according to a label reliability evaluation criterion, we generate the selection indicators $\boldsymbol{s}$ in order to obtain the candidates set $\mathcal{S}$ via Eq. (5.3). In the model update step, the set $\mathcal{S}$ along with the initial labeled set $\mathcal{L}$ is regarded as the new training set $\mathcal{D}$, *i.e.*, $\mathcal{D} = \mathcal{L} \cup \mathcal{S}$. The set $\mathcal{D}$ will be utilized to re-train the model so as to make the model more robust. During training iterations, the candidates set $\mathcal{S}$ in each step is enlarged continuously. In this way, we can progressively learn a more stable model.

To be specific, for our progressive strategy EUG, we adopt an end-to-end CNN model with temporal average pooling (**ETAP-Net**) as the feature embedding function ϕ. The ETAP-Net is an adaption of ResNet-50 architecture for video inputs, where we add a fully-connected layer and a temporal average pooling layer before the classification layer. As shown in Figure 6.2, for each tracklet, all frames are processed to obtain frame-level feature embedding. The frame features within a tracklet are then element-wise averaged as the tracklet feature representation by the temporal average pooling layer. In the label estimation step, for each unlabeled video tracklet, the pseudo label is assigned by the identity label of its nearest labeled neighbor in the tracklet feature space. The distance between them is considered as the dissimilarity cost, which is used to measure the reliability of its pseudo label.

5.3.3 Progressive and Effective Sampling Strategy

It is crucial to obtain the appropriately selected candidates $\mathcal{S}$ to exploit the unlabeled data. In this procedure, two significant aspects are mainly considered: First, how to ensure the reliability of selected pseudo-labeled samples? Second, what is an effective sampling criterion on the unlabeled data for one-example person re-ID?

Discussion on Sampling Strategy. The reliability of pseudo labels originates from two main challenges in the one-example learning setting. (1) the initial labeled data are too few to depict the detailed underlying distribution. (2) learning a CNN model on a not-yet-reliable training set may not improve the re-ID performance. The interplay of these two factors hinders the further performance improving. Therefore, it is irrational to incorporate excessive pseudo-labeled data into training at the initial iteration.

Discussion on Sampling Criterion. The previous works sample the unlabeled data from confident to uncertain ones according to the classification loss. However, the loss from classification prediction does not well fit the retrieval evaluation. Moreover, it is far away to train a robust identity classifier in the one-example setting, where each class has only one sample for training. The classifier may easily over-fit the one-example labeled data and may not learn the intrinsic distinction in classification. Therefore, the classification prediction may be not reliable on an unseen sample.

Our step-wise Solution. To address aforementioned two problems, we propose (1) a dynamic sampling scheme, which progressively increases the number of selected pseudo-labeled samples; (2) an effective sampling criterion, which takes the distance in the feature space as a measure of reliability.

The proposed dynamic sampling scheme steadily increases the size of selected candidates set $|\mathcal{S}|$ during iterations. It starts with a small proportion of pseudo-labeled data at the beginning stages, and then incorporates more diverse samples in the following stages. As the training iteration goes, the reliability of pseudo labels grows steadily, because the re-ID model becomes more robust and discriminative. Therefore, more pseudo-labeled candidates can be adopted into training.

For sampling criterion, instead of classification prediction, we adopt the Nearest Neighbors (NN) classifier for the label estimation. For the one-example setting, the NN classifier in the feature space may be a better choice, since similar input data always have similar feature representations. The NN classifier assigned the label of

Algorithm 1: Exploit the Unknown Gradually

```
Require: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$, enlarging factor $p \in (0,1)$, initialized
    CNN model $\boldsymbol{\theta}_0$.
Ensure: The best CNN model $\boldsymbol{\theta}^*$.
Initialize the selected pseudo-labeled data $\mathcal{S}_0 \leftarrow \emptyset$, sampling size $m_1 \leftarrow p \cdot n_u$,
    iteration step $t \leftarrow 0$, best validation performance $V^* \leftarrow 0$
while $m_{t+1} \leq |\mathcal{U}|$ do
    $t \leftarrow t+1$
    Update training set: $\mathcal{D}_t \leftarrow \mathcal{L} \cup \mathcal{S}_{t-1}$
    Train the CNN model $(\boldsymbol{\theta}_t, \boldsymbol{w}_t)$ based on $\mathcal{D}_t$.
    Generate the selection indicators $\boldsymbol{s}_t$ via Eq. (6.9)
    Update $\mathcal{S}_t$ based on $\boldsymbol{s}_t$ via Eq. (5.3)
    Update the sampling number: $m_{t+1} \leftarrow m_t + p \cdot n_u$
end while
for i $\leftarrow$ 1 to T do
    Evaluate $\boldsymbol{\theta}_i$ on the validation set $\rightarrow$ performance $V_i$
    if $V_i > V^*$ then
        $V^*, \boldsymbol{\theta}^* \leftarrow V_i, \boldsymbol{\theta}_i$
    end if
end for
```

each unlabeled data by its nearest labeled neighbor in feature space. We define the confidence of label estimation as the distance between the unlabeled data and its nearest labeled neighbor. For the candidates selection, we select some of top reliable pseudo-labeled data according to their label estimation confidence.

More formally, we define the **dissimilarity cost** for each unlabeled data $x_i \in \mathcal{U}$ as:

$$d(\boldsymbol{\theta}; x_i) = \min_{x_l \in \mathcal{L}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_l)||_2, \tag{5.4}$$

The cost is the minimum l_2 distance between the unlabeled data x_i and an arbitrary labeled data $x_l \in \mathcal{L}$ in the feature space parameterized by $\boldsymbol{\theta}$. The dissimilarity cost is considered as the criterion for measuring the confidence of pseudo-labeled data. For the candidates selection, at the iteration step t, we sample the pseudo-labeled candidates into training by setting the selection indicator $\boldsymbol{s}_t$ as follows:

$$\boldsymbol{s}_t = \arg \min_{||s||_0 = m_t} \sum_{i=n_l+1}^{n_l+n_u} s_i d(\boldsymbol{\theta}; x_i), \tag{5.5}$$

where the m_t denotes the size of selected pseudo-labeled set. As the iteration step t increases, we enlarge the size of sampled pseudo-labeled data by set $m_t = m_{t-1} + p \cdot n_u$. $p \in (0,1)$ is the **enlarging factor** which indicates the speed of enlarging the candidates set during iterations. Eq. (6.9) selects the top m_t nearest unlabeled data for all the labeled data at the iteration step t. As described in Algorithm 2, we evaluate the model $\phi(\boldsymbol{\theta}_t; \cdot)$ on the validation set at each iteration step and output the best model. In the one-example experiment, we take another video-based person re-ID training set as the validation set.

How to find a proper enlarging factor $\boldsymbol{p}$? An *aggressive* choice is to set p to a very large value, which urges m_t to increase rapidly. As a result, the sampled pseudo-labeled candidates may not be reliable enough to train a robust CNN model. A *conservative* option is to set p to a very small value, which means m_t progressively enlarges with a small change in each step. This option tends to result in a very stable increase in the performance and a promising performance in the end. The disadvantage is that it may require an excessive number of stages to touch great performance.

5.4 Experiments

5.4.1 Settings and Implementation Details

Experiment Setting. For one-example experiments, we use the same protocol as [66]. In both datasets, we randomly choose one tracklet in camera 1 for each identity as initialization. If there is no tracklet recorded by camera 1 for one identity, we randomly select one tracklet in the next camera to make sure each identity has one video tracklet for initialization. Note that as discussed in Section 6.2, [66, 129] are using the same one-example setting in experiments.

Implementation Details. We use PyTorch [78] for all experiments. As discussed in Section 5.3.2, we take ETAP-Net as our basic CNN model for training on video-based re-ID. In experiments, we take ImageNet [41] pre-trained ResNet-50 model with last classification layer removed as the initialization of ETAP-Net. For training as a classification task for each identity, an additional fully-connected layer with batch normalization [32] and a classification layer are appended at the end of the model. The parameters of the first three residual blocks of ResNet-50 are kept fixed in training to save GPU memory and boost iterations. In training, we randomly sample 16 frames as the input for each tracklet. In label estimation and evaluation steps, all the frames are processed by the CNN model to get the representations for each tracklet, which

Methods	rank-1	rank-5	rank-20	mAP
Baseline (one-example)	36.16	50.20	61.86	15.45
DGM+IDE[129]	36.81	54.01	68.51	16.87
Stepwise[66]	41.21	55.55	66.76	19.65
EUG ($p = 0.30$)	42.77	56.51	67.17	21.12
EUG ($p = 0.20$)	48.68	63.38	72.57	26.55
EUG ($p = 0.15$)	52.32	64.29	73.08	29.56
EUG ($p = 0.10$)	57.62	69.64	78.08	34.68
EUG ($p = 0.05$)	**62.67**	**74.94**	**82.57**	**42.45**
Baseline (supervised)	80.75	92.07	96.11	67.39

Table 5.1 Comparison with the state-of-the-art methods on MARS. All the methods are conducted based on the same backbone model ETAP-Net. Baseline (one-example) is the initial model trained on one-example labeled data. p is the enlarging factor that indicates the enlarging speed of the sampled subset. At the bottom we provide the Baseline (supervised) result as a upper bound where 100% training data are labeled.

are further l_2 normalized and used to calculate the Euclidean distance. We adopt the stochastic gradient descent (SGD) with momentum 0.5 and weight decay 0.0005 to optimize the parameters for 70 epochs with batch size 16 in each iteration. The overall learning rate is initialized to 0.1 and changed to 0.01 in the last 15 epochs.

5.4.2 Comparison with the State-of-the-Art Methods

We compare our method to DGM [129] and step-wise [66] on the one-example task. Note that although [66, 129] claim them as *unsupervised* methods, they are actually *one-example* methods in experiments, because they require at least one labeled tracklet for each identity. Since the performances of both works were reported based on hand-crafted features, to make a fair comparison, we reproduce their methods using the same backbone model ETAP-Net (ResNet-50) as ours. The re-ID performance on MARS and DukeMTMC-VideoRe-ID are summarized in Table 5.1 and Table 5.2. On the MARS dataset, we achieve surprising result with rank-1 accuracy 62.67%, mAP 42.45% with enlarging factor 0.05, which greatly outperform the state-of-the-art result by 21.46 points and 22.8 points (absolute), respectively. The great performance gap between [66, 129] and ours is due to the excessive not-yet-reliable pseudo-labeled data incorporated at the first iteration. The estimation errors are accumulated during iterations and thus limit the further enhancement.

Methods	rank-1	rank-5	rank-20	mAP
Baseline (one-example)	39.60	56.84	66.95	33.27
DGM+IDE[129]	42.36	57.92	69.31	33.62
Stepwise[66]	56.26	70.37	79.20	46.76
EUG ($p = 0.30$)	63.82	78.64	87.04	54.57
EUG ($p = 0.20$)	68.95	81.05	89.46	59.50
EUG ($p = 0.15$)	69.08	81.19	88.88	59.21
EUG ($p = 0.10$)	70.79	83.61	89.60	61.76
EUG ($p = 0.05$)	**72.79**	**84.18**	**91.45**	**63.23**
Baseline (supervised)	83.62	94.59	97.58	78.34

Table 5.2 Comparison with the state-of-the-art methods on DukeMTMC-VideoReID. All the methods are conducted based on the same backbone model ETAP-Net. Baseline (one-example) is the initial model trained on one-example labeled data. p is the enlarging factor that indicates the enlarging speed of the sampled subset. At the bottom we provide the Baseline (supervised) result as a upper bound where 100% training data are labeled.

Moreover, Baseline (one-example) and Baseline (supervised) are our initial model and the upper bound model, respectively. Baseline (one-example) takes only the one-example labeled data as the training set and do not exploit the unlabeled data. Baseline (supervised) is conducted on the fully supervised setting that all tracklets in the dataset are labeled and adopted in training. Specifically, we achieve 26.51 points and 33.19 points rank-1 improvements over the Baseline (one-example) on MARS and DukeMTMC-VideoReID, respectively.

5.4.3 Algorithm Analysis

Analysis on the sampling criteria.

As mentioned in Section 5.3.3, some previous works such as SPL take the classification loss as the criterion. The label estimation and evaluation performances of sampling by classification loss and by dissimilarity cost are illustrated in Figure 5.3 and Table 5.3. From the figure, we observe the huge performance gaps for both label estimation and evaluation. The label estimations of both criteria achieve similar and high precision at the beginning stage. However, the label estimation accuracy gap between two criteria gradually enlarges. As a result, the performance of the classification loss criterion is only enhanced to a limited extent and drops quickly in the subsequence. Table 5.3 shows the evaluation performance differences of the two criteria with different enlarging factors. With the same enlarging factor, the criterion of sampling by dissimilarity cost

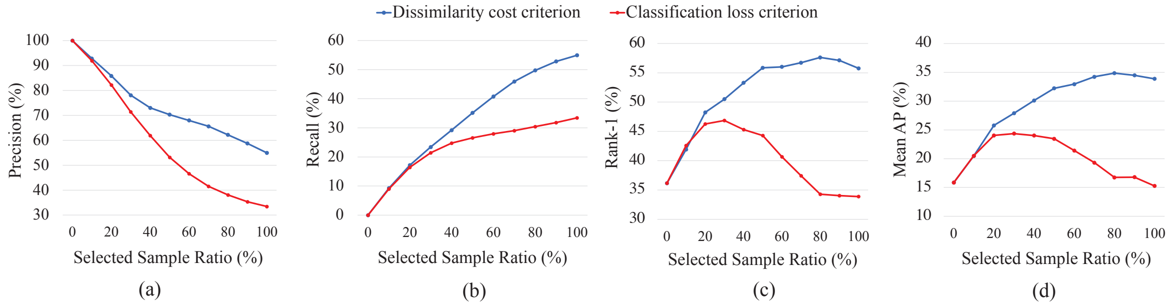


Fig. 5.3 Comparison with two sampling criteria on MARS when the Enlarging Factor $p = 0.1$. (a) and (b): Precision and recall of the pseudo label prediction of selected pseudo-labeled candidates during iterations with different sampling criteria. (c) and (d): Rank-1 accuracy and mAP of person re-ID on the evaluation set during iterations with different sampling criteria. The x-axis stands for the percentage of selected data from entire unlabeled data for updating. Each solid point indicates an iteration step.

Enlarging Factor	Criteria	rank-1	rank-5	mAP
$p = 0.05$	Classification	48.33	62.67	25.35
	Dissimilarity	62.67	74.94	42.45
$p = 0.10$	Classification	46.86	60.25	24.23
	Dissimilarity	57.62	69.64	34.68
$p = 0.15$	Classification	46.53	60.12	24.03
	Dissimilarity	52.32	64.29	29.56
$p = 0.20$	Classification	45.91	59.95	23.56
	Dissimilarity	48.68	63.38	26.55
$p = 0.30$	Classification	41.86	56.01	20.24
	Dissimilarity	42.77	56.51	21.12

Table 5.3 Comparison of the two criteria on MARS. The "Classification" and "Dissimilarity" denotes the EUG methods with the classification loss criterion and the dissimilarity cost criterion, respectively. Note for that with the same enlarging factors, the dissimilarity cost criterion always lead to a superior performance.

always leads to the superior performance. When the enlarging factor is set to 0.05, the best rank-1 accuracy on evaluation for classification loss and dissimilarity cost is 48.33% and 62.67%, respectively.

Analysis over iterations. Figure 5.4 illustrates the label estimation performance and evaluation performance over iterations. At the initial iteration, the precision of pseudo label for the selected subset (blue line) is relatively high, since EUG only adopts a few of the most reliable samples. In later stages, as EUG gradually incorporates more difficult and diverse samples, the precision drops along with the recall (red

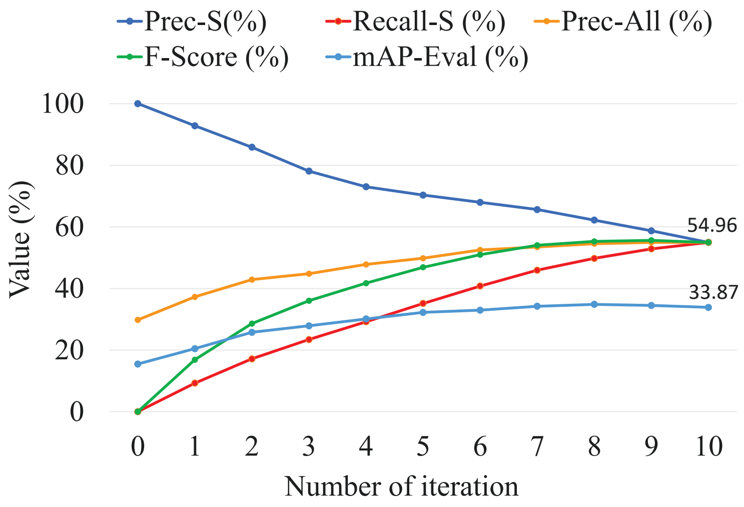


Fig. 5.4 The label estimation performance with the enlarging factor = 0.1 over iterations on MARS. "Prec-S", "Recall-S" and "F-Score" denote the label estimation precision, recall and F-score for the *selected* pseudo-labeled candidates. "Prec-All" denotes the overall label estimation precision for *all* the unlabeled data. "mAP-Eval" represents the mAP performance of the evaluation on the test set. Note that on all the unlabeled data the overall label estimation accuracy is constantly increasing, which indicates the model learns much information throughout iterations.

line) rising. In spite of the descending of precision, the F-score of label estimation (green line) continuous increases. Throughout iterations, the precision of pseudo label estimation for all the unlabeled data (orange line) constantly increases from 29.8% to 54.96%, which indicates the model grows robust steadily. At the last few iterations, the evaluation performance stops to increase, because the gain of adding new samples is offset by the loss of excessive pseudo label errors.

Analysis on the enlarging factor. For the iteration t, $t * p$ percent of unlabeled tracklets with reliable pseudo labels are sampled for updating the model. The effectiveness of enlarging factor p is shown in Figure 5.5. Two conclusions can be inferred: First, the model always achieves a better performance if we enlarge the selected set at a slower speed. The huge gaps among the five curves show that the great impact of the enlarging factor. Second, we observe that the gaps among the five curves are relatively small in the first several iterations and gradually enlarge in the later iterations. It shows the estimation errors are accumulated during iterations. This is because that the performance of the trained CNN model highly depends on the reliability of the training set. As a result, the evaluation performances appear obvious different in the last few iterations.

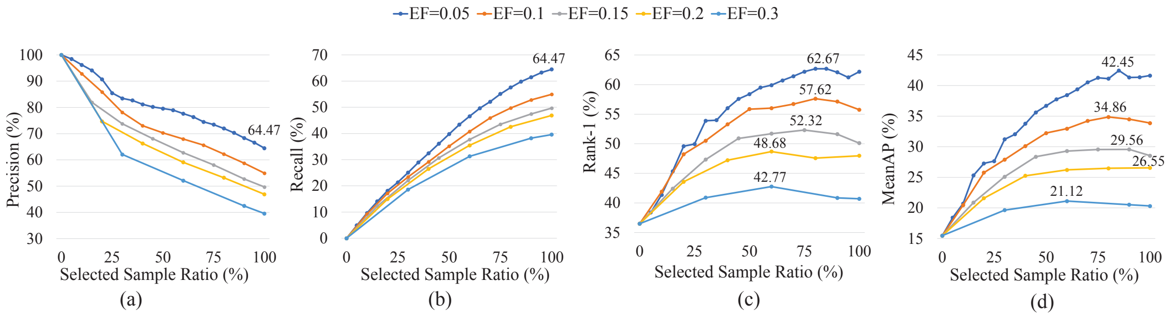


Fig. 5.5 Comparison with different value of enlarging factor on MARS. (a) and (b) : Precision and recall of the pseudo label prediction of selected candidates with different enlarging factors. (c) and (d) : Rank-1 and mAP of person re-ID on the evaluation set with different enlarging factors. "EF" denotes the enlarging factor. The x-axis stands for the ratio of selected data from entire unlabeled data for updating. Each solid point indicates an iteration step. Note for that the lower enlarging factor is beneficial for improving performance.

5.4.4 Visualization

We visualize the selected samples for an identity during iterations in Figure 5.6. Since the initial tracklets is captured from the side view of the pedestrian, the two unlabeled tracklets captured from the same side are easily selected in iteration 0. In iteration 1 and 2, some tracklets in the behind or front view of the pedestrian are selected. The above tracklets are relatively easier for sampling. Further, in iteration 5 and 6, video tracklets suffering from obstructing and color variance are sampled. In iteration 7, samples with pedestrian of small size and dark background are selected. It's clear that the samples are selected from easy to hard, from similar to diverse. Note that there is no tracklet selected for this identity in iteration 3 and 4, which indicates the huge difficulty gap. There are also four mismatches in iteration 5, 6, and 7, in which the pedestrian is very similar to the ground truth identity, with the same pink shirt, gray pants, and long hair.

5.5 Conclusion

Label estimation for unlabeled tracklets is crucial for one-example person re-ID. The challenge in the one-example setting is that the pseudo labels are not reliable enough, which prevents the trained model from improving robust. To solve this problem, we propose a dynamic sampling strategy to start with easy and reliable unlabeled

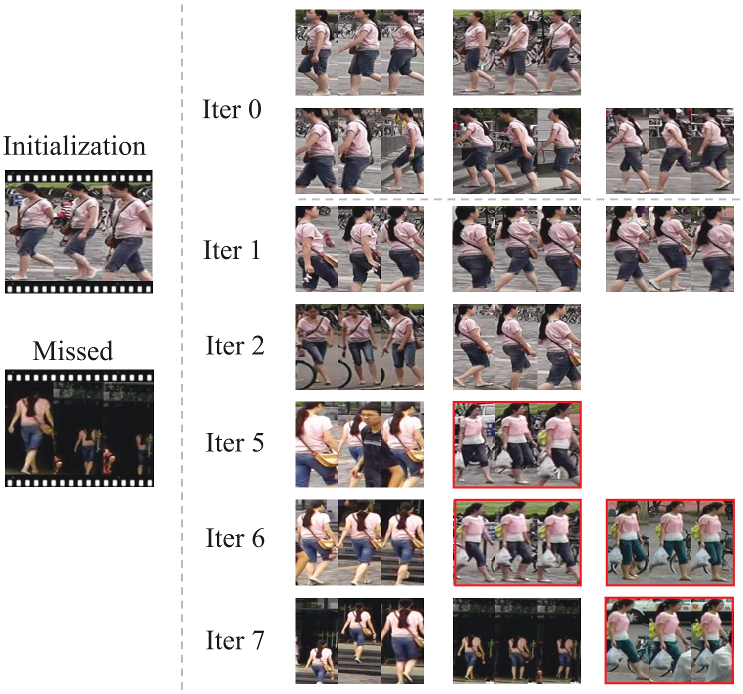


Fig. 5.6 The selected pseudo-labeled tracklets for an identity example on MARS with the enlarging factor $p = 0.1$. Error estimated samples are in red rectangles. All the tracklets incorporated in the former iterations are naturally selected by later ones. For this identity, one tracklet is missed, and four false samples are selected. Observe that the tracklet selected is easy and reliable at the beginning stage and difficult and diverse in the later stage.

samples and gradually incorporating diverse tracklets for updating the model. We found that if we enlarge the selected set at a slower speed, the model achieves a better performance. In addition, we present a sampling criterion to remarkably improving the performance of label estimation. Our method surpasses the state-of-the-art method by 21.46 points (absolute) in rank-1 accuracy on MARS, and 16.53 points (absolute) on DukeMTMC-VideoReID. In sum, the proposed method is effective in exploiting the unlabeled data and reducing the annotation work load for one-example video-based person re-ID.

Chapter 6

Progressive Learning for Person Re-Identification with One Example

In the previous chapter, we propose a progressive model for video-based re-ID on the one-example setting, where *each identity has one labeled example and abundant unlabeled examples.* While in this chapter, we improve the progressive framework by gradually exploiting the unlabeled data.

The framework also contains two process that updated iteratively: (1) train the Convolutional Neural Network (CNN) model and (2) estimate pseudo labels for the unlabeled data. We split the training data into three parts, *i.e.*, labeled data, pseudo-labeled data, and unselected data. Initially, the re-ID model is trained using the labeled data. For the subsequent model training, we update the CNN model by the joint training on the three data parts. The proposed joint training method can optimize the model by both the data with labels (or pseudo labels) and the data without any reliable labels. For the label estimation step, instead of using a static sampling strategy, we propose a progressive sampling strategy to increase the number of the selected pseudo-labeled candidates step by step. We select a few candidates with most reliable pseudo labels from unlabeled examples as the pseudo-labeled data, and keep the rest as unselected data. During iterations, the unselected unlabeled data are dynamically transferred to be selected pseudo-labeled data.

Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 21.6 points (absolute, *i.e.*, 62.8% vs. 41.2%) on MARS, and 16.6 points on DukeMTMC-VideoReID. Extended to the few-example setting, our approach with only 20% labeled data surprisingly achieves comparable performance to the supervised state-of-the-art method with 100% labeled data.

6.1 Introduction

Person re-identification (re-ID) aims at spotting the person-of-interest from non-overlapping camera views[113, 114]. In recent years, deep convolutional neural networks (CNN) has led to impressive successes in the field of re-ID [29, 116, 142, 147]. Most existing re-ID methods, in particular deep learning models, adopt the supervised learning approach. These methods rely on the full annotations, *i.e.*, the identity labels of all the training data from multiple cross-view cameras. However, it is impractical to annotate large-scale data due to the dramatic cost for the human annotator to identify pedestrians. The exhaustive manual efforts hence limit the applications where many different cameras exist in practice.

Recently, semi-supervised learning methods for person re-ID [3, 66, 129] are of particular interest. This work mainly focuses on the one-example setting, in which only one example is required to be labeled for each identity. Annotating one example for each identity is extremely easy compared to the exhaustive manual efforts on cross-camera labelling. In the one-example setting, the key challenge is how to utilize the abundant unlabeled data effectively. A typical approach for the one-example task is to first generate the pseudo labels for the unlabeled data. Then the initial labeled data and some selected pseudo-labeled data are considered as an enlarged training set. Lastly, this new training set is used to train the re-ID model.

Most existing methods employ a static strategy to determine the quantity of selected pseudo-labeled data for further training. For example, Fan *et al.* [16] and Ye *et al.* [129] compare the prediction confidences of pseudo-labeled samples with a *pre-defined* threshold. The samples with confidence higher than the fixed threshold are then selected for the subsequent training. During iterations, these algorithms select a fixed size of pseudo-labeled training data from beginning to end. However, in the initial stage, only a few label predictions are reliable due to the very few labeled examples. As the iteration goes, the model gets more robust, resulting in more accurate label predictions. Therefore, keeping the size of the selected data fixed would hinder the further improvement of the performance.

We propose a progressive learning framework to better exploit the unlabeled data for person re-ID with limited exemplars. Initially, a CNN model is trained on the one-example labeled samples. We then iteratively update the CNN by two steps, the label estimation step and the model update step. In the first step, we generate the pseudo labels for all unlabeled samples, and select some reliable pseudo-labeled data for training according to the prediction confidence. Different from existing methods [66, 129], the selected subset is continuously enlarged during iterations according to a sampling strategy. In the second step, same as the labeled data, the pseudo labels of selected examples are incorporated in the model updating. Particularly, we start with a small-size subset of pseudo-labeled samples, which only includes the most reliable and easiest ones. In the subsequent stages, it gradually selects a growing number of pseudo-labeled data to incorporate more difficult and diverse data.

In our preliminary version [114], we intensively investigate two significant aspects for one-example video-based re-ID, *i.e.*, the progressive sampling strategy and the dissimilarity sampling criterion. The proposed method achieves surprisingly good performance for the one-example video-based person re-ID where each identity has a labeled tracklet (62 image frames on average). In this chapter, we expand it to a more difficult task, the one-example image-based re-ID, where only one image instead of a tracklet is labeled for each identity. We found it hard to obtain a good initial model on the very few labeled images. It inspires us to further improve our approach by exploiting the abundant unselected unlabeled data in an self-supervised manner. These data account for a large proportion of training examples at initial iteration stages but have been overlooked in all the previous methods. Although we cannot obtain reliable pseudo identity labels for these unlabeled data, these images depict rich details of persons.

In this chapter, we split training data into three parts, labeled data, selected data (pseudo-labeled) and unselected data (unlabeled). We then propose a joint learning method to simultaneously train the CNN model on three data splits. Specifically, as shown in Figure 6.1, there are three different data sources. For the labeled data and selected pseudo-labeled data, we apply an identity classifier on their CNN features and further optimize the model by comparing the identity predictions and the (pseudo-) labels. For those unselected unlabeled data, we use the exclusive loss to optimize the model without any labels. The classification loss pulls representations of the same identity data close to each other, while the exclusive loss pushes representations of all the unselected samples away from each other. We also extend our approach from one-example setting to the few-example setting. The experiments show that our method can reduce 80% annotation cost with only a 4.3% rank-1 accuracy drop on MARS.

Our contributions are summarized as follows:

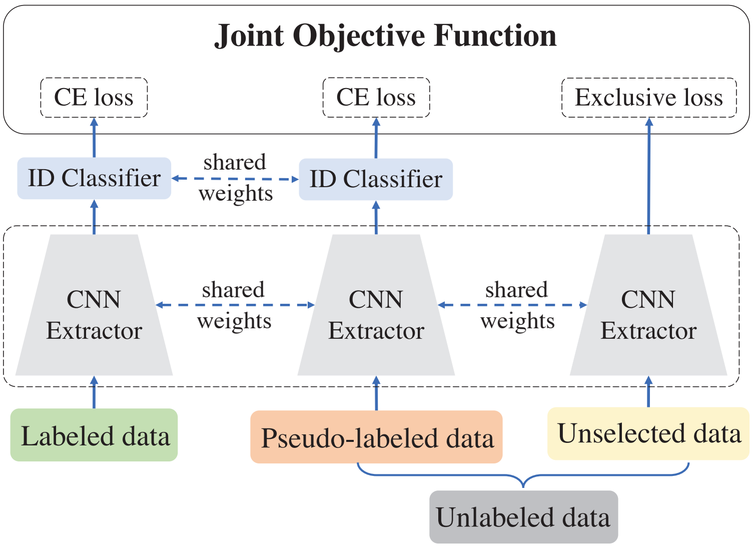


Fig. 6.1 The proposed joint training procedure for few-example person re-ID. "CE loss" denotes the Cross-Entropy loss. We select a few unlabeled data with reliable pseudo labels as the pseudo-labeled data. On the labeled and pseudo-labeled data, we optimize the CNN model by an ID classifier and corresponding Cross Entropy loss. For the unselected data where no reliable pseudo label is available, an exclusive loss is utilized to learn a discriminative feature embedding from images without any labels.

- We propose a progressive method for few-example person re-ID to better exploit the unlabeled data. This method adopts a dynamic sampling strategy to uncover the unlabeled data. We start with reliable samples and gradually include diverse ones, which significantly makes the model robust.

- We apply a distance-based sampling criterion for label estimation and apply candidates selection to remarkably improve the performance of label estimation.

- We propose a joint learning method to simultaneously train the CNN model on the labeled, pseudo-labeled and unselected unlabeled data.

- Our method achieves surprisingly superior performance on the one-example setting, outperforming the state-of-the-art by 21.6 points (absolute) on MARS and 16.6 points (absolute) on DukeMTMC-VideoReID.

- Our approach can be readily extended into the few-example setting (with 20% labeled data). It achieves comparable rank-1 performance (76.46%) compared to the state-of-the-art performance (79.80%) in the supervised setting (with 100% labeled data).

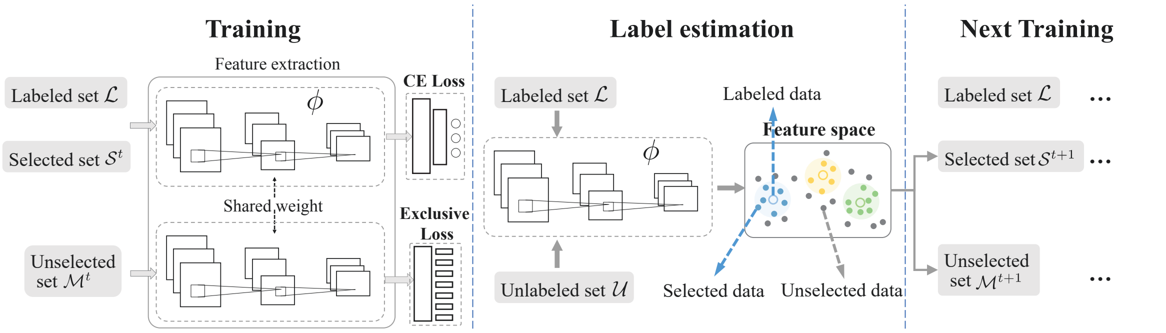

Training
Feature extraction
Labeled set $\mathcal{L}$
Selected set $\mathcal{S}^t$
ϕ
Shared weight
Unselected set $\mathcal{M}^t$
CE Loss
Exclusive Loss
Label estimation
Labeled set $\mathcal{L}$
ϕ
Unlabeled set $\mathcal{U}$
Labeled data
Feature space
Selected data
Unselected data
Next Training
Labeled set $\mathcal{L}$ ...
Selected set $\mathcal{S}^{t+1}$ ...
Unselected set $\mathcal{M}^{t+1}$ ...

Fig. 6.2 Overview of the proposed iterative framework. In each iteration, (1) we train the CNN model by the joint learning on the *labeled* data, *selected pseudo-labeled* data, and *unselected unlabeled* data. We utilize the labeled data and the selected pseudo-labeled data by an identity classification learning with their (pseudo-) labels. On the unselected dataset where no reliable pseudo label is available, we apply the exclusive loss directly on the images to optimize the model without any identity label. (2) In the label estimation step, we select a few reliable pseudo-labeled candidates from unlabeled data $\mathcal{U}$ to the selected set $\mathcal{S}$ according to the distance in feature space. Nodes with different colors in the feature space box denote different identity samples.

6.2 Related Works

Semi-supervised learning. Semi-supervised learning [37, 45, 82] takes advantages from both labeled and unlabeled data to solve the given task. Some semi-supervised approaches [38] use graph representations in recent years. Kipf *et al.* [38] encode the graph structure directly using a neural network model and train on a supervised target for all nodes with labels. Most recently, with the great success of the Generative Adversarial Network (GAN) [22], many researchers adopt semi-supervised learning to explore images generated by GAN [87, 147]. Salimans *et al.* [87] present a variety of new architectural features and training procedures that apply to the generative adversarial networks framework.

6.3 The Progressive Model

We first introduce the framework overview of the proposed method in Section 6.3.1, and the preliminaries in Section 6.3.2 Then we illustrate the two key parts of our method, *i.e.*, the joint learning method in Section 6.3.3 and the label estimation in Section 6.3.4. Lastly, we present the overall progressive iteration strategy in Section 6.3.5.

6.3.1 Framework Overview

In this work, we propose a progressive learning method to exploit the unlabeled data gradually and steadily. In general, as shown in Figure 6.2, our method updates the model by the following two steps iteratively: 1. train the CNN model by the joint learning on the *labeled* data, *selected pseudo-labeled* data, and *unselected unlabeled* data; 2. select a few reliable pseudo-labeled candidates from unlabeled data according to a prediction reliability criterion. Specifically, in the first iteration, all the unlabeled data are unselected. During iterations, we continuously enlarge the set of selected candidates to progressively learn a robust and stable model. In the label estimation step, the pseudo label is assigned to the candidate by the identity label of its nearest labeled neighbor in the training data feature space. The distance between them is considered as the dissimilarity cost, which is the measure of reliability for the pseudo label.

6.3.2 Preliminaries

We first introduce the necessary notations for the one-example re-ID task. Let x and y denote the pedestrian visual data and the identity label, respectively. The visual data x can be either a person image for image-based re-ID, or a tracklet (a series of person images) for video-based re-ID. For the training in the one-example re-ID task, we have the labeled data set $\mathcal{L} = \{(x_1, y_1), ..., (x_{n_l}, y_{n_l})\}$ and the unlabeled data set $\mathcal{U} = \{x_{n_l+1}, ..., x_{n_l+n_u}\}$. Usually, these data are utilized in an identity classification way to train the re-ID model $\phi(\boldsymbol{\theta}, \cdot)$. For the evaluation stage, the trained CNN model ϕ is used to embed both query data and gallery data into the feature space. The query result is the ranking list of all gallery data according to the Euclidean Distance between the query data and each gallery data, *i.e.*., $||\phi(\boldsymbol{\theta}; x_q) - \phi(\boldsymbol{\theta}; x_g)||$, where x_q and x_g denote the query data and the gallery data, respectively. To exploit abundant unlabeled data, we predict the pseudo label $\hat{y}_i$ for each unlabeled data $x_i \in \mathcal{U}$ and select a few reliable ones for the identity classification learning. Let $\mathcal{S}^t$ and $\mathcal{M}^t$ denote the selected pseudo-labeled dataset and unselected unlabeled data set at t-th step, respectively.

6.3.3 The Joint Learning Method

We first introduce the model updating step. At the t-th iteration, we have three kinds of the data source for model training, *i.e.*, the labeled data $\mathcal{L}$, the selected pseudo-labeled data $\mathcal{S}^t$ and the remaining unselected data $\mathcal{M}^t$. We utilize the labeled data $\mathcal{L}$ and the pseudo-labeled data $\mathcal{S}^t$ by an identity classification learning with their (pseudo-) labels. For the unselected data, their pseudo labels are not-yet-reliable and may harm the model training. Therefore, we utilize the exclusive loss to optimize the CNN model on the unselected set.

The Exclusive Loss aims at learning a discriminative embedding on $\mathcal{M}^t$ without any identity labels. In general, we optimize the CNN model by learning to distinguish samples, rather than identities. To push each unselected data $x_i \in \mathcal{M}^t$ away from the other data $x_j \in \mathcal{M}^t, i \neq j$ in the feature space, we define the following target for unsupervised feature learning:

$$\max_{\boldsymbol{\theta}} \sum_{\substack{x_i, x_j \in \mathcal{M}^t \\ x_i \neq x_j}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_j)||, \tag{6.1}$$

where $||\cdot||$ denotes the Euclidean distance between two embedded features.

To solve Eq. (6.1) in an efficient way, we have the following approximation. Let $v_i = \tilde{\phi}(\boldsymbol{\theta}; x_i)$ be the L2-normalized feature embedding for the data x_i, *i.e.*, $||v_i|| = 1$. Since $||v_i - v_j||^2 = 2 - v_i^{\mathrm{T}} v_j$, maximizing the Euclidean distance between data x_i and x_j is equivalent to minimize the cosine similarity $v_i^{\mathrm{T}} v_j$. Therefore, Eq. (6.1) can be optimized by a softmax-like loss:

$$\ell_e(\boldsymbol{V}; \tilde{\phi}(\boldsymbol{\theta}; x_i)) = -\log \frac{\exp({v_i}^{\mathrm{T}} \tilde{\phi}(\boldsymbol{\theta}; x_i)/\tau)}{\sum_{j=1}^{|\mathcal{M}^t|} \exp({v_j}^{\mathrm{T}} \tilde{\phi}(\boldsymbol{\theta}; x_i)/\tau)}, \tag{6.2}$$

where $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{M}^t| \times n_\phi}$ is a lookup table that stores the features of each unselected data x_i. τ is a temperature parameter that controls the concentration level of the distribution. A higher temperature τ leads to a softer probability distribution. Inspired by [117], we adopt the lookup table $\boldsymbol{V}$ to avoid the exhaustive computation of extracting features from all data at each training step. In the forward operation, we compute cosine similarities between data x_i and all the other data by $\boldsymbol{V}^T \tilde{\phi}(\boldsymbol{\theta}; x_i)$. During backward, we update the i-th column of the table $\boldsymbol{V}$ by $v_i \leftarrow \frac{1}{2}(v_i + \tilde{\phi}(\boldsymbol{\theta}; x_i))$ and then L2-normalize v_i to a unit vector.

The exclusive loss is a self-supervised auxiliary loss to learn discriminative representations from the unlabeled data. It can also be viewed as a regularization term, as it introduces the remaining unlabeled data to the model training process and thus avoids simply overfitting to the labeled data and selected pseudo-labeled data. During the model optimization, to achieve the target that any two unlabelled samples are repelled away, the exclusive loss forces the model to learn to distinguish the difference of the input images (or tracklets). Therefore, the learned representation is expected to focus more on details of an input identity. During this procedure, the model could access more samples during training. Although these samples have neither ground truth labels nor the reliable pseudo labels, they can still provide some weak supervision information by exploiting the differences between human images.

Joint objective Function. There are three data parts, *i.e.*, the labeled data, pseudo-labeled data, and the remaining unselected unlabeled data. We jointly optimize our model on these three data parts. On the labeled dataset $\mathcal{L}$ which we have the ground truth identity labels, we follow recent works [16, 56, 149] to train the re-ID model. We have the following objective function for model training:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=1}^{n_l} \ell_{\texttt{CE}}(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),y_i), \tag{6.3}$$

where $f(\boldsymbol{w};\cdot)$ is an identity classifier, parameterized by $\boldsymbol{w}$, to classify the embedded feature $\phi(\boldsymbol{\theta};x_i) \in \mathbb{R}^{n_\phi}$ into a k-dimension confidence estimation, in which k is the number of *identities.* $\ell_{\texttt{CE}}$ denotes the cross entropy loss on the identity label prediction $f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)) \in \mathbb{R}^k$ and its ground truth identity label y_i. Similarly, we can optimize the model on the pseudo-labeled data set $\mathcal{S}$ by

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \sum_{i=n_l+1}^{n_l+n_u} s_i \ell_{\texttt{CE}}(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),\hat{y}_i), \tag{6.4}$$

where $s_i \in \{0,1\}$ is the selection indicator for the unlabeled sample x_i, which is generated from previous label estimation step. s_i determines whether we should select pseudo-labeled data $(x_i,\hat{y}_i)$ for identity classification training. We will discuss it later in Section 6.3.4.

Considering the three data splits, we design the following objective function for the model training at t-th iteration:

$$\begin{aligned} \min_{\boldsymbol{\theta},\boldsymbol{w}} \lambda \sum_{i=1}^{n_l} & \ell_{\texttt{CE}}(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),y_i)+ \\ \lambda \sum_{i=n_l+1}^{n_l+n_u} & s_i^{t-1} \ell_{\texttt{CE}}(f(\boldsymbol{w};\phi(\boldsymbol{\theta};x_i)),\hat{y}_i)+ \\ (1-\lambda) \sum_{i=n_l+1}^{n_l+n_u} & (1-s_i^{t-1}) \ell_e(\boldsymbol{V};\tilde{\phi}(\boldsymbol{\theta};x_i)), \end{aligned} \tag{6.5}$$

where λ is a hyper-parameter to adjust the contribution of the identity classification loss $\ell_{\texttt{CE}}$ and the exclusive loss ℓ_e. The Eq. (6.5) consists of three loss parts. The first part is the ID classification loss on the labeled set $\mathcal{L}$. The second one is the ID classification loss on the selected pseudo-labeled set $\mathcal{S}^t$. The last one is the exclusive loss on the unselected set $\mathcal{M}^t$.

6.3.4 The Effective Sampling Criterion

We then introduce the label estimation step. It is crucial to obtain the appropriately selected candidates $\mathcal{S}^t$ to exploit the unlabeled data. The previous works sample the unlabeled data from confident to uncertain ones according to the classification loss [14]. However, the loss from classification prediction does not well fit the retrieval evaluation. Moreover, the classifier may easily over-fit the one-example labeled data. Thus it may be not robust in predicting identity. To address this problem, we propose an effective sampling criterion, which takes the distance in the feature space as a measure of pseudo label reliability.

For the label estimation on unlabeled data, we adopt the Nearest Neighbors (NN) classifier instead of classification prediction. The NN classifier assigns the pseudo label for each unlabeled data by its nearest labeled neighbor in the feature space. We define the confidence of label estimation as the distance between the unlabeled data and its nearest labeled neighbor. For the candidates selection, we select a few top reliable pseudo-labeled data according to their label estimation confidence.

More formally, we estimate the pseudo label for each unlabeled data $x_i \in \mathcal{U}$ by:

$$x^*, y^* = \arg\min_{(x_l, y_l) \in \mathcal{L}} ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x_l)||, \tag{6.6}$$

$$d(\boldsymbol{\theta}; x_i) = ||\phi(\boldsymbol{\theta}; x_i) - \phi(\boldsymbol{\theta}; x^*)||, \tag{6.7}$$

$$\hat{y}_i = y^*, \tag{6.8}$$

where $d(\boldsymbol{\theta}; x_i)$ is the **dissimilarity cost** of label estimation. The dissimilarity cost is used as the criterion for measuring the confidence of pseudo-labeled data. To select candidates, at the iteration step t, we sample the pseudo-labeled candidates into training by setting the selection indicators as follows:

$$\boldsymbol{s}^t = \arg\min_{||\boldsymbol{s}^t||_0 = m_t} \sum_{i=n_l+1}^{n_l+n_u} s_i d(\boldsymbol{\theta}; x_i), \tag{6.9}$$

where m_t denotes the size of selected pseudo-labeled set and $\boldsymbol{s}^t$ is the vertical concatenation of all s_i. Eq. (6.9) selects the top m_t nearest unlabeled data for all the labeled data at the iteration step t.

6.3.5 The overall Iteration Strategy

We iteratively train the CNN model and then estimate labels for unlabeled data. At each iteration, we first optimize the model by Eq. (6.5). Then we estimate labels for unlabeled data by Eq. (6.8) and select some reliable ones by applying the trained model on Eq. (6.9). Since the initial labeled data are too few to depict the detailed underlying distribution, it is irrational to incorporate excessive pseudo-labeled data in training at the initial iteration.

We propose a dynamic sampling strategy to ensure the reliability of selected pseudo-labeled samples. It starts with a small proportion of pseudo-labeled data at the beginning stages and then incorporates more diverse samples in the following stages. We start our framework by setting $m_0 = 0$ and $\mathcal{M}^0 = \mathcal{U}$, *i.e.*, optimizing the model by (1) identity classification training on labeled data $\mathcal{L}$ and (2) unsupervised training by exclusive loss on the unlabeled data. In the later iterations, we progressively increase the size of selected pseudo-labeled candidates set $|\mathcal{S}^t|$. At iteration step t, we enlarge the size of sampled pseudo-labeled data by set $m_t = m_{t-1} + p \cdot n_u$, where $p \in (0, 1)$ is the **enlarging factor** which indicates the speed of enlarging the candidates set during iterations. As described in Algorithm 2, we evaluate the model $\phi(\boldsymbol{\theta}; \cdot)$ on the validation set at each iteration step and output the best model. In the one-example experiment, we take another person re-ID training set as the validation set.

How to find a proper enlarging factor $\boldsymbol{p}$ for real-life applications? The enlarging factor controls the speed of enlarging the reliable pseudo-labeled candidates set during iterations. Smaller enlarging factor indicates lower enlarging speed, therefore, more iteration steps and training time. In the real-life application, this factor is a trade-off between efficiency and accuracy. An *aggressive* choice is to set p to a very large value, which urges m_t to increase rapidly. As a result, the sampled pseudo-labeled candidates may not be reliable enough to train a robust CNN model. A *conservative* option is to set p to a very small value, which means m_t progressively enlarges with a small change in each step. This option tends to result in a very stable increase in the performance and a promising performance in the end. The disadvantage is that it may require an excessive number of stages to touch great performance.

Algorithm 2: The proposed framework

```
Require: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$, enlarging factor $p \in (0,1)$, initialized CNN model $\boldsymbol{\theta}_0$.
Ensure: The best CNN model $\boldsymbol{\theta}^*$.
Initialize the selected pseudo-labeled data $\mathcal{S}_0 \leftarrow \emptyset$, sampling size $m_1 \leftarrow p \cdot n_u$, iteration step $t \leftarrow 0$, best validation performance $V^* \leftarrow 0$.
while $m_{t+1} \leq |\mathcal{U}|$ do
    $t \leftarrow t+1$
    Update the model $(\boldsymbol{\theta}_t, \boldsymbol{w}_t)$ on $\mathcal{L}$, $\mathcal{S}^t$ and $\mathcal{M}^t$ via Eq. (6.5).
    Estimate pseudo labels for $\mathcal{U}$ via Eq. (6.8)
    Generate the selection indicators $\boldsymbol{s}_t$ via Eq. (6.9)
    Update the sampling size: $m_{t+1} \leftarrow m_t + p \cdot n_u$
end while
for i $\leftarrow$ 1 to T do
    Evaluate $\boldsymbol{\theta}_i$ on the validation set $\rightarrow$ performance $V_i$
    if $V_i > V^*$ then
        $V^*, \boldsymbol{\theta}^* \leftarrow V_i, \boldsymbol{\theta}_i$
    end if
end for
```

6.4 Experimental Analysis

6.4.1 Settings and Implementation Details

Experiment Setting. For one-example experiments, we use the same protocol as [66]. In all datasets, we randomly choose an image/trackelt from camera 1 for each identity as initialization. If there is no data recorded by camera 1 for one identity, we randomly select a sample from the next camera to make sure each identity has one sample for initialization. Note that [66, 129] are using the same one-example setting in experiments.

Implementation Details.

We adopt ResNet-50 with the last classification layer removed as our feature embedding model ϕ to conduct all the experiments. We initialize it by the ImageNet [41] pre-trained model. To optimize the model by the (pseudo-) label loss, we append an additional fully-connected layer with batch normalization and a classification layer on the top of the CNN feature extractor. For the exclusive loss, we process the unlabeled feature by a fully-connected layer with batch normalization, followed by the L2-normalization operation. Following [117], the temperature scalar τ in Eq. (6.2) is set to 0.1. We set λ in Eq. (6.5) to be 0.8 for all the experiments. In each model

Table 6.1 Comparison with the state-of-the-art methods on two image-based large-scale re-ID datasets. Baseline (one-example) is the initial model trained on one-example labeled data. Baseline (supervised) shows the upper bound performance where 100% training data are labeled. p is the enlarging factor that indicates the enlarging speed of the selected pseudo-labeled subset.

Settings	Method	rank-1	rank-5	rank-10	rank-20	mAP
Supervised	Baseline [149]	83.1	92.5	95.0	96.9	63.7
One-Example	Baseline [149]	26.0	41.4	49.2	59.6	9.0
	Ours ($p = 0.30$)	35.5	52.8	60.5	68.6	13.4
	Ours ($p = 0.20$)	41.4	59.6	66.4	73.5	17.4
	Ours ($p = 0.15$)	44.8	61.8	69.1	76.1	19.2
	Ours ($p = 0.10$)	51.5	66.8	73.6	79.6	23.2
	Ours ($p = 0.05$)	**55.8**	**72.3**	**78.4**	**83.5**	**26.2**

Table 6.2 Comparison with the state-of-the-art methods on DukeMTMC-reID. Baseline (one-example) is the initial model trained on one-example labeled data. Baseline (supervised) shows the upper bound performance where 100% training data are labeled. p is the enlarging factor that indicates the enlarging speed of the selected pseudo-labeled subset.

Settings	Methods	rank-1	rank-5	rank-10	rank-20	mAP
Supervised	Baseline [149]	71.0	83.2	87.3	89.9	49.3
One-Example	Baseline [149]	16.4	27.9	32.8	39.0	6.8
	Ours ($p = 0.30$)	23.3	35.7	42.2	48.0	11.1
	Ours ($p = 0.20$)	30.0	43.4	49.2	54.8	15.1
	Ours ($p = 0.15$)	35.1	49.1	54.3	60.0	18.2
	Ours ($p = 0.10$)	40.5	53.9	60.2	65.5	21.8
	Ours ($p = 0.05$)	**48.8**	**63.4**	**68.4**	**73.1**	**28.5**

updating step, the stochastic gradient descent (SGD) with momentum 0.5 and weight decay 0.0005 is used to optimize the parameters for 70 epochs with batch size 16. The overall learning rate is initialized to 0.1. In the last 15 epochs, to stabilize the model training, we change the learning rate to 0.01 and set $\lambda = 1$. For the experiments on video-based re-ID datasets, we simply add a temporal average pooling layer on the CNN extractor, where we element-wisely average features of all frame within a tracklet.

Table 6.3 Comparison with the state-of-the-art methods on MARS. Baseline (one-example) is the initial model trained on one-example labeled data. Baseline (supervised) shows the upper bound performance where 100% training data are labeled. p is the enlarging factor that indicates the enlarging speed of the selected pseudo-labeled subset.

Settings	Methods	rank-1	rank-5	rank-10	rank-20	mAP
Supervised	Baseline [149]	80.8	92.1	94.6	96.1	63.7
One-Example	Baseline [149]	36.2	50.2	57.2	61.9	15.5
	DGM+IDE [129]	36.8	54.0	59.6	68.5	16.9
	Stepwise [66]	41.2	55.6	62.2	66.8	19.7
	Ours ($p = 0.30$)	44.5	58.7	65.7	70.6	22.1
	Ours ($p = 0.20$)	49.6	64.5	69.8	74.4	27.2
	Ours ($p = 0.15$)	52.7	66.3	71.9	76.4	29.9
	Ours ($p = 0.10$)	57.9	70.3	75.2	79.3	34.9
	Ours ($p = 0.05$)	**62.8**	**75.2**	**80.4**	**83.8**	**42.6**

Table 6.4 Comparison with the state-of-the-art methods on DukeMTMC-VideoReID. Baseline (one-example) is the initial model trained on one-example labeled data. Baseline (supervised) shows the upper bound performance where 100% training data are labeled. p is the enlarging factor that indicates the enlarging speed of the selected pseudo-labeled subset.

Settings	Methods	rank-1	rank-5	rank-10	rank-20	mAP
Supervised	Baseline [149]	83.6	94.6	96.9	97.6	78.3
One-Example	Baseline [149]	39.6	56.8	62.5	67.0	33.3
	DGM+IDE [129]	42.4	57.9	63.8	69.3	33.6
	Stepwise [66]	56.3	70.4	74.6	79.2	46.8
	Ours ($p = 0.30$)	66.1	79.8	84.9	88.3	56.3
	Ours ($p = 0.20$)	69.1	81.2	85.6	89.6	59.6
	Ours ($p = 0.15$)	69.3	81.4	85.9	89.2	59.5
	Ours ($p = 0.10$)	71.0	83.8	87.4	90.3	61.9
	Ours ($p = 0.05$)	**72.9**	**84.3**	**88.3**	**91.4**	**63.3**

6.4.2 Comparison with the State-of-the-Art Methods

There are two recent works designed for one-example video-based person re-ID, *i.e.*, DGM [129] and Stepwise [66]. Note that although [66, 129] claim them as *unsupervised* methods, they are actually *one-example* methods in experiments, because they require at least one labeled tracklet for each identity. We compare our method to them on the one-example video-based re-ID task. Since the performances of both works were reported based on hand-crafted features, to make a fair comparison, we reproduce their

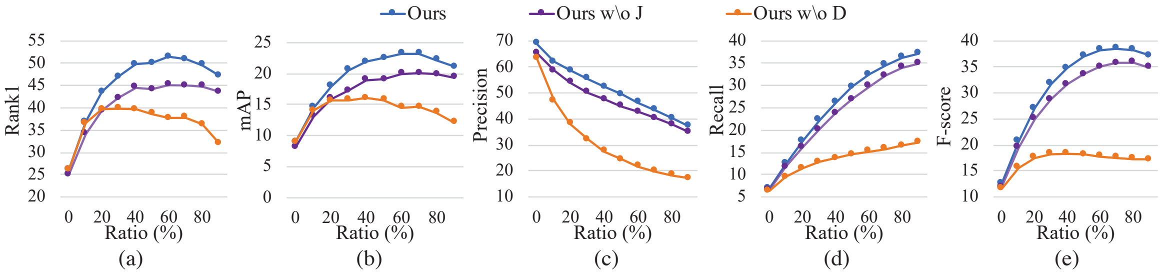


Fig. 6.3 Ablation studies on Market-1501. We validate the effectiveness of the two parts of our method, *i.e.*, the joint learning method (denoted as J) and the dissimilarity cost (denoted as D). The enlarging factor p is set to 0.1 in the comparison. (a) and (b): Rank-1 accuracy and mAP on the evaluation set during iterations. (c), (d) and (e): Precision, recall, and F-score of the label prediction of selected pseudo-labeled candidates during iterations. The x-axis stands for the percentage of selected data over entire unlabeled data. Each solid point indicates an iteration step.

methods using the same backbone model ResNet-50 as ours. The re-ID performance of our method on the four large-scale re-ID datasets are summarized in Table 7.1, Table 7.2, Table 7.3 and Table 7.4. With only one labeled example for each identity, our method achieves surprising performance on both image-based and video-based re-ID task.

Moreover, we compare our method to two baseline methods, *i.e.*, the Baseline (one-example) and Baseline (supervised), which are our initial model and the upper bound model (100% data are labeled), respectively. Baseline (one-example) takes only the one-example labeled data as the training set and do not exploit the unlabeled data. Baseline (supervised) is conducted on the fully supervised setting that all data are labeled and adopted in training. Specifically, we achieve 29.8, 32.4, 26.6 and 33.3 points of rank-1 accuracy improvement over the Baseline (one-example) on Market-1501,DukeMTMC-reID, MARS and DukeMTMC-VideoReID, respectively. The superior performances on the four large-scale datasets validate the effectiveness of our proposed method.

6.4.3 Ablation Studies

To validate the effectiveness of each component in our proposed method, we conduct ablation studies on the two key parts of our methods, *i.e.*, the joint learning method and the dissimilarity criterion, as shown in Table 6.5 and Figure 6.3. All experiments share the same training parameters and initial labeled images.

Table 6.5 Ablation studies w.r.t. rank-1 and mAP on Market-1501 (image-based) and MARS (video-based). "Ours w/o D" denotes the model without the dissimilarity cost, *i.e.*, using classification loss as the criterion. Note that "Ours w/o D" is optimized on all the three data parts. "Ours w/o J" indicates the model without the joint learning method, *i.e.*, only optimized by the identity classification training on labeled and selected pseudo-labeled data. "Ours" is the full model.

Enlarging factor	Methods	Market-1501		MARS	
		rank-1	mAP	rank-1	mAP
$p = 0.30$	Ours w/o D	35.2	13.2	42.0	20.3
	Ours w/o J	28.9	10.5	42.8	21.1
	Ours	35.5	13.4	44.5	22.1
$p = 0.20$	Ours w/o D	36.5	13.7	45.5	23.5
	Ours w/o J	36.2	14.0	48.7	26.6
	Ours	41.4	17.4	49.6	27.2
$p = 0.10$	Ours w/o D	39.8	16.1	46.4	24.1
	Ours w/o J	45.1	20.1	57.6	34.7
	Ours	51.5	23.2	57.9	34.9
$p = 0.05$	Ours w/o D	40.3	16.2	48.1	25.2
	Ours w/o J	49.8	22.5	62.6	42.4
	Ours	55.8	26.2	62.8	42.6

The effectiveness of the joint learning method. We compare our method to the model trained without the joint learning method, denoted as "Ours w/o J" in Table 6.5. The "Ours w/o J" model is only optimized by the identity classification loss on the labeled and selected pseudo-labeled data, as proposed in the preliminary version [114]. The comparison results on the two datasets prove the effectiveness of the joint learning method. Compared to the great improvement on the image-based task (Market-1501), the improvement of the video-based task (MARS) is relatively small. The main reason is that the one-example initial model in the video-based re-ID task is much more robust compared to the image-based one, since an initial tracklet contains many images (62 frames on average on MARS) of the same identity. It can be seen from the accuracy difference of the initial label predictions on all the unlabeled data, *i.e.*, 30.0% on MARS while 11.9% on Market-1501. Exploiting the unlabeled data with a relatively robust model may not benefit the feature learning a lot.

The effectiveness of the sampling criteria. As mentioned in Section 6.3.4, some previous works such as SPL take the classification loss as the criterion. The label estimation accuracy and re-ID performances of sampling by classification loss and by dissimilarity cost are illustrated in Figure 6.3 and Table 6.5. We observe the huge performance gaps in Figure 6.3 for both label estimation and evaluation. The label estimations of both criteria achieve similar and high precision at the beginning stage. However, the label estimation accuracy gap between two criteria gradually enlarges. As a result, the performance of the classification loss criterion is only enhanced to a limited extent and drops quickly in the subsequence. Table 6.5 shows the evaluation performance differences between the two criteria with different enlarging factors. "Ours w/o D" denotes the method with classification loss as the criterion. With the same enlarging factor, the criterion of sampling by dissimilarity cost always leads to superior performance.

6.4.4 Algorithm Analysis

Analysis over iterations.

Figure 6.3 illustrates the label estimation performance and re-ID performance over iterations. Since we only collect a few most reliable unlabeled samples as pseudo-labeled data, the precision score of label estimation is relatively high at the beginning. As iteration goes, we adopt more unlabeled data into the pseudo-labeled set, resulting in a continuous precision drop of the label estimation. However, in this procedure, the recall score of label estimation gradually increases as more correctly estimated pseudo-labeled data are used. The overall label estimation performance, the F-score, appears a rapid increase at the first several iterations and a slight performance drop in the last iterations. Interestingly, the re-ID evaluation performances, *i.e.*, Rank-1 and mAP scores, show a similar curve with F-score, which indicates the label estimation quality is the key factor in the One-Example task.

Analysis on the enlarging factor.

The enlarging factor p is a key parameter in our framework. It controls the speed of enlarging pseudo-labeled candidates set during iterations. Smaller enlarging factor indicates lower enlarging speed, therefore, more iteration steps and training time. The results of different enlarging factors can be found in Figure 6.4. As we can see, in experiments, a smaller enlarging factor always yield a better performance. It is consistent with human intuition since each enlarging step is more cautious and thus the label estimation is more accurate. We could also find that the gaps among the five curves are relatively small in the first several iterations and gradually enlarge in the later iterations, which shows the estimation errors are accumulated during iterations.

Qualitative Analysis.

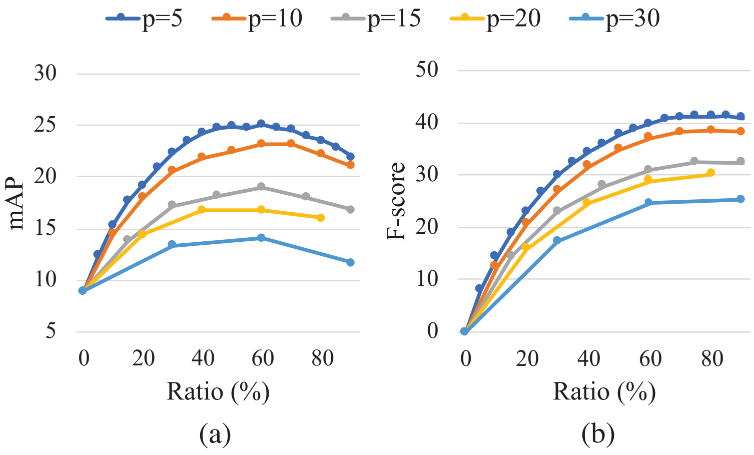


Fig. 6.4 Comparison with the different value of enlarging factor p on Market-1501. (a) mAP of person re-ID on the evaluation set with different enlarging factors. (b) F-score of the label prediction of selected candidates with different enlarging factors. The x-axis stands for the ratio of selected data from the entire unlabeled set. Each solid point indicates an iteration step.

We visualize our selected pseudo-labeled samples for an identity during iterations in Figure 6.5. As shown in the left, the initial labeled sample is captured from the side view of the pedestrian, wearing a white shirt and black pants. At the beginning iteration stages (iteration 1 to 4), the selected samples are very similar samples that are also captured from the side view of the pedestrian. In iteration 5, samples in the behind view of the pedestrian are selected into the pseudo-labeled set. The above samples are relatively easy for the model to distinguish. In iteration 6, there is no sample selected for this identity, indicating there is a difficulty gap here in the pseudo label generation. Further, in iteration 6 and 7, some samples from other pedestrians are selected for this identity by mistake, indicated by the red box in the figure. Although these are error cases, the selected data are very similar to the initial labeled sample that they share almost the same cloth appearance. In the last two stages, our method selects the difficult samples which are suffering from both color variance and view change (captured from the back view). It's clear that the samples are selected from easy to hard, from similar to diverse. There are also four samples missed for this identity, *i.e.*, assigned to other identities by mistake. However, the person in these images is either in small size or surrounded by the dark background.

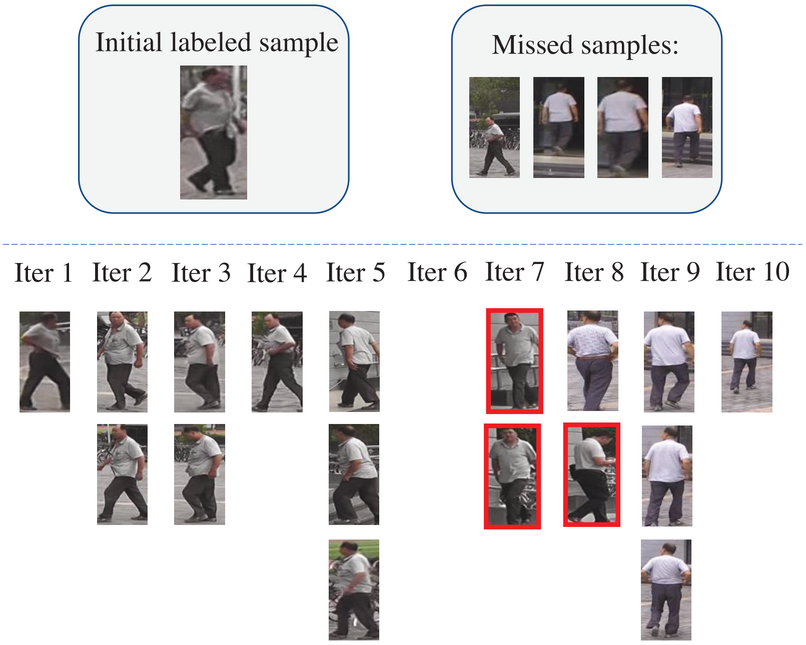


Fig. 6.5 The selected pseudo-labeled samples for an identity example on Market-1501. We use the enlarging factor p of 0.1. Error estimated samples are in red rectangles. All the samples selected in the former iterations are naturally selected again by later ones. We only show the newly added samples of each iteration in the figure. For this identity, four samples are missed, and three false samples are selected. As shown in the figure, the selected samples are easy and reliable at the beginning stage and difficult and diverse in the later stage.

6.4.5 Evaluation on the Few-example Setting

Our method can be easily extended to the few-example re-ID task by annotating more labeled data for initialization. We report the few-example performances on the Market-1501 dataset (see Table 6.7) and the MARS dataset (see Table 6.6). The performances of our method in different ratios of labeled data are reported. On the Marker-1501 dataset, our method outperforms the state-of-the-art method SPACO [70] by a large margin. On the MARS dataset, when using 20% labeled training data, our method achieves 76.5% rank-1 and 60.3% mAP, which is close to the state-of-the-art supervised methods with 100% labeled tracklets (upper bound). Although it costs more annotation effort for the few-example task comparing to the one-example task, it can easily achieve the competitive results to supervised performance with only a part of training data labeled.

Table 6.6 Comparison to the state-of-the-art on MARS. The listed works are fully supervised methods, and the other performance we reported are in the few-example setting. The number in the bracket indicates the percentage of used labeled training data.

Settings	Methods	rank-1	rank-5	rank-20	mAP
Supervised	MSCAN[48]	71.8	86.6	93.1	56.0
	K-reciprocal[149]	73.9	-	-	68.4
	IDTriplet[29]	79.8	91.4	-	67.7
Semi-supervised	Ours (10%)	72.2	84.8	91.8	54.2
	Ours (20%)	76.5	88.4	93.3	60.3
	Ours (40%)	79.2	91.1	95.6	65.5
	Ours (60%)	80.6	91.6	95.7	66.8

Table 6.7 Comparison to the state-of-the-art on Market-1501. The number in the bracket indicates the percentage of used labeled training data.

Settings	Methods	rank-1	rank-5	rank-20	mAP
Supervised	IDE [142]	72.5	-	-	46.0
	K-reciprocal[149]	77.1	-	-	63.6
	Siamese [146]	79.5	90.9	-	59.9
	GAN [147]	83.9	-	-	66.0
	IDTriplet[29]	84.9	94.2	-	69.1
Semi-supervised	SPACO (20%) [70]	68.3	-	-	-
	Ours (5%)	70.1	84.2	92.1	43.6
	Ours (10%)	80.7	90.4	95.8	58.3
	Ours (20%)	82.5	92.4	97.2	63.6

6.5 Conclusion

To tackle the one-example re-ID task, we propose a progressive training framework and the joint training method. In the framework, we iteratively train the CNN model and estimate pseudo labels for the unlabeled data. For the label estimation step, we propose a progressive sampling strategy to enlarge the pseudo-labeled data set. For the model training, our proposed joint training method can effectively exploit the labeled data, the selected pseudo-labeled data, and the unselected unlabeled data. Our method outperforms the baseline by 29.8 points (absolute) in rank-1 accuracy on Market-1501, and surpasses the state-of-the-art method by 21.6 points (absolute) on MARS. The promising performance improvement demonstrates the effectiveness of our method.

Chapter 7

Unsupervised Person Re-identification

Most person re-identification (re-ID) approaches are based on supervised learning, which requires intensive manual annotation for training data. However, it is not only resource-intensive to acquire identity annotation but also impractical to label the large-scale real-world data. To relieve this problem, in this chapter we propose a bottom-up clustering (BUC) approach to jointly optimize a convolutional neural network (CNN) and the relationship among the individual samples. Our algorithm considers two fundamental facts in the re-ID task, *i.e.*, ***diversity*** across different identities and ***similarity*** within the same identity. Specifically, our algorithm starts with regarding individual sample as a different identity, which maximizes the diversity over each identity. Then it gradually groups similar samples into one identity, which increases the similarity within each identity. We utilizes a diversity regularization term in the bottom-up clustering procedure to balance the data volume of each cluster. Finally, the model achieves an effective trade-off between the *diversity* and *similarity*. We conduct extensive experiments on the large-scale image and video re-ID datasets, including Market-1501, DukeMTMC-reID, MARS and DukeMTMC-VideoReID. The experimental results demonstrate that our algorithm is not only superior to state-of-the-art unsupervised re-ID approaches, but also performs favorably than competing transfer learning and semi-supervised learning methods.

7.1 Introduction

Person re-identification (re-ID) aims at matching a target person in a set of query pedestrian images. In recent years, the widespread adoption of deep convolutional

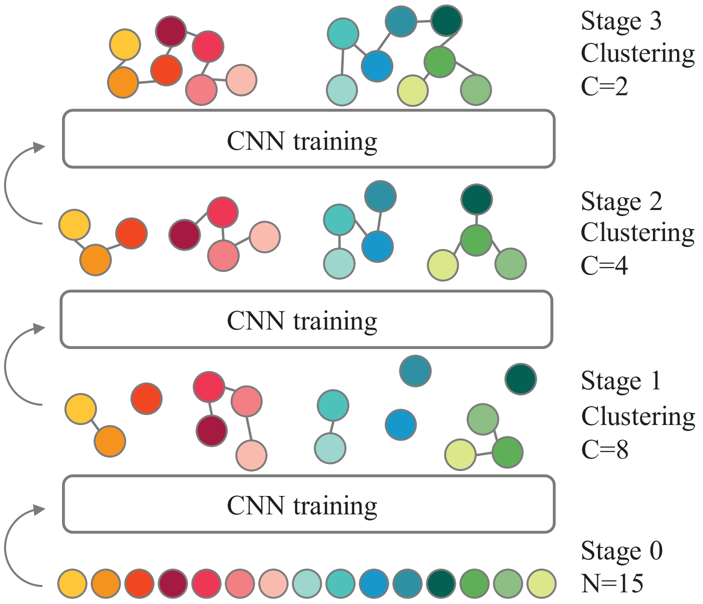


Fig. 7.1 The bottom-up clustering method. Each circle denotes an individual image for training. N denotes the number of training samples, while C denotes the number of clusters after clustering. After the current stage of network training, we apply clustering based on the previous clustering result and the feature similarity of the current stage. By clustering from bottom to up, individual pedestrian samples are gathered to represent an identity.

neural networks (CNN) has led to impressive progress in the field of re-ID [51, 100, 130]. However, supervised re-ID methods require intensive manual labeling. It is expensive and not applicable to the real-world applications. The limited generalization ability motivates the research into unsupervised approaches for person re-ID.

Traditional unsupervised methods focus on hand-crafted features [17, 54, 57], salience analysis [103, 139] and dictionary learning [39]. These methods produce much lower performance than supervised methods and are not applicable to large-scale real-world data. In recent years, some transfer learning methods [11, 28, 80] are proposed upon the success of deep learning [53]. These methods usually learn an identity-discriminative feature embedding on the source dataset, and transfer the learned features to the unseen target domain. However, these methods require a large amount of annotated source data, which cannot be regarded as pure unsupervised approaches.

Previous deep learning based "unsupervised" person re-ID approaches leverage the prior knowledge learned from other re-ID datasets. However, we aim to solve the problem in a more challenge and practical setting, *i.e.*, without any re-ID annotation.

To learn discriminate features in this difficult condition, we propose a novel Bottom-Up Clustering method (**BUC**) for unsupervised re-ID that maximizes the diversity over the identities while maintaining the similarity within each identity. As illustrated in Fig. 7.1, during the training process, the individual samples are gathered into clusters, and the clusters will merge gradually. Specifically, our framework BUC applies network training and the bottom-up clustering in an iterative way. We propose the repelled loss that can optimize the network without actually having any label and obtain decent initial accuracy. At the beginning of network training, we view individual images as exemplars, *i.e.*, each image belongs to a distinct cluster. We then gradually incorporate similarity within identities by a bottom-up clustering, which is to merge similar images (clusters) into one cluster. Moreover, in practice, different identities should have a similar probability to be captured by cameras, and thus the image number for different clusters should be balanced. To enforce this property, we incorporate a diversity regularization term in the merging procedure. Finally, during the bottom to up clustering procedure, our framework exploits the similarity and the diversity to learn discriminative features.

Our contributions can be summarized in four-fold:

- We propose a bottom-up clustering framework to solve the unsupervised re-ID problem. By exploiting the intrinsic diversity among identities and similarity within each identity, our framework can learn robust and discriminative features.

- We adopt the repelled loss to optimize the model without labels. The repelled loss directly optimizes the cosine distance among each individual sample / cluster. It can facilitate the model to exploit the similarity within each cluster and maximize the diversity among each identity.

- We propose a diversity regularization term to enable the balanced image number in each cluster. It makes the clutering results align with the real world distribution.

- The experimental results demonstrate that our approach is superior to the state-of-the-art methods on both image-based and video-based re-ID datasets. We achieves top-1 accuracy of 66.2% on Market-1501 [142] and 61.1% on MARS [141]. Moreover, the one-shot re-ID methods utilize more annotation than ours, whereas our approach also obtains a higher performance than them.

7.2 Related Work

Most re-ID methods are in a supervised manner, in which sufficient labeled person pairs across cameras are given. They mainly focus on designing feature representations [140] or learning robust distance metrics [54, 145]. Recently, deep learning methods achieve great success [51, 100, 142, 146] by simultaneously learning the image representations and similarities. In this chapter, we focus on the unsupervised setting and do not discuss more supervised methods here.

Unsupervised person re-identification. The existing fully unsupervised methods usually fall into three categories, designing hand-craft features [17, 54, 57], exploiting localized salience statistics [103, 139] or dictionary learning based methods [39, 123]. However, it is a challenging task to design suitable features for images captured by different cameras, under different illumination and view condition. These methods are unable to explicitly exploit the cross-view discriminative information without pairwise identity labels. Thus the performance of these methods is much weaker than supervised methods. Recently, Xiao *et al.* [117] propose the OIM loss for semi-supervised person search. It can also be used for unsupervised re-ID. Compared to OIM, our BUC has three advantages. (1) We constrain the feature to distribute on a unit sphere to improve its robustness. (2) We design the cluster merging to exploit the similarity among identities. (3) We propose a diversity regularization term to avoid the model collapse.

Another category of the unsupervised method makes use of additional information [11, 52, 80]. Recently, cross-domain transfer learning is used in the unsupervised re-ID task [28, 105], where information from an external source dataset is utilized. Fan *et al.* [28] propose a progressive method, where the K-means clustering and the IDE [142] network pre-trained on the source dataset are updated iteratively. Wang *et al.* [105] propose to learn an attribute-semantic and identity discriminative representation from the source dataset, which is transferable to the target domain. There are also some recent works [66, 127, 129] focusing on the unsupervised video-based re-ID. However, these methods require some very useful annotations of the dataset, *i.e.*, the total number of identities and their appearance. To conduct experiments, they annotate each identity with a labeled video tracklet, which only reduces part of the annotation workload. As discussed in [114], these approaches are actually the one-example methods. Different from these methods, our work focuses on the fully unsupervised setting in which there is ***no annotation*** on the dataset.

Unsupervised feature learning. Unsupervised feature learning is widely studied in many tasks, such as image recognition, image classification, and image retrieval [97].

Some works use hand-crafted features combined with conventional clustering methods [24, 25, 90]. However, the hand-designed features are not as effective as deeply learned features. A number of works [4, 15] sample patches from images and generate labels for the patches as supervision. In [15], exemplar-CNN is proposed to discriminate among a set of surrogate classes, where the surrogate classes are formed by applying a variety of transformations to randomly sampled image patches.

Wu *et al.* [115] propose a non-parametric softmax classifier and use noise-contrastive estimation to tackle the computational challenges. Different from these works, our BUC not only considers the diversity over each sample but also exploits the similarity within each class. Comparing with these unsupervised feature learning methods on the classification task, our BUC obtains superior performance.

7.3 Methodology

7.3.1 Preliminary

Given a training set $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ of N images, our goal is to learn a feature embedding function $\phi(\theta; x_i)$ from $\mathbf{X}$ without any manual annotation, where parameters of ϕ are collectively denoted as θ. This feature embedding function can be applied to the testing set, $\mathbf{X}^t = \{x_1^t, x_2^t, ...x_{N_t}^t\}$ of N_t images, and the query set $\mathbf{X}^q = \{x_1^q, x_2^q, ...x_{N_q}^q\}$ of N_q images. During the evaluation, we use the feature of a query image $\phi(\boldsymbol{\theta}; x_i^q)$ to search the similar image features from the testing set. The query result is a ranking list of all testing images according to the Euclidean distance between the feature embedding of the query and testing data, *i.e.*, $d(x_i^q, x_i^t) = \|\phi(\boldsymbol{\theta}; x_i^q) - \phi(\boldsymbol{\theta}; x_i^t)\|$. The feature embeddings are supposed to assign a higher rank to similar images and keep the images of a different person a low rank.

To learn the feature embedding, traditional methods usually learn the parameters with manual annotations. That is, each image x_i is associated with a label y_i, where $1 \leq y_i \leq k$ and k is the number of identities. A classifier $f(\boldsymbol{w}; \phi(\theta; x_i)) \in \mathbb{R}^k$ parameterized by $\boldsymbol{w}$ is used to predict the identity of the image x_i. The classifier parameter $\boldsymbol{w}$ and the embedding parameter θ are jointly optimized by the following objective function:

$$\min_{\boldsymbol{\theta}, \boldsymbol{w}} \sum_{i=1}^{N} \ell(f(\boldsymbol{w}; \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{7.1}$$

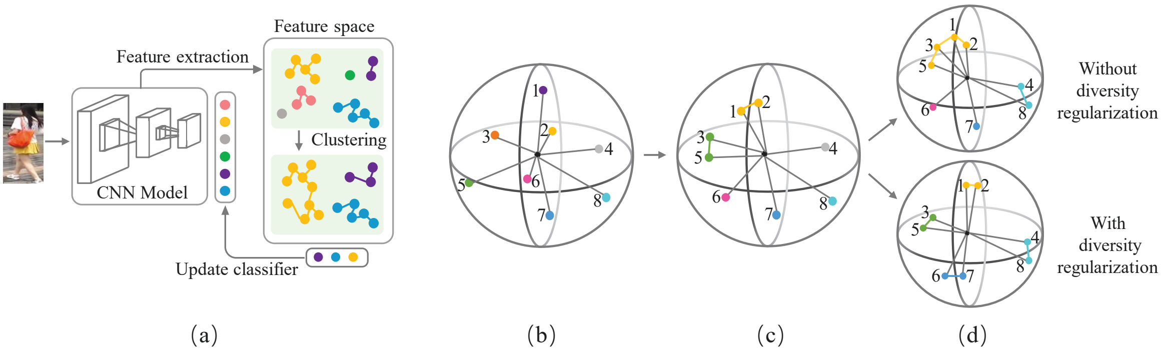


Fig. 7.2 (a). The proposed framework takes unlabeled images as input to train the network and extract the image features for clustering. The framework do three steps alternatively, *i.e.*, extracts the CNN feature for each image, merges clusters over the whole training set, and re-train the CNN model. Fig. (b)-(d) depict the cluster merging procedure. In this example, each step we merge two cluster pairs with minimum dissimilarity. The solid points with the same color represent images in the same cluster. The colored line indicates the connected two clusters have been merged into one. In (b), the learned features discriminatively span a unit sphere, which the diversity is maximized. In (c), after merging the clusters, feature embeddings of the same cluster get closer in the sphere. In (d), the upper sphere shows the cluster merging result without diversity regularization: (Point 1, Point 3) and (Point 4, Point 8) have the shortest distances, and are then merged into one cluster. The lower sphere shows the cluster merging result with diversity regularization: though the distance between the yellow and green clusters is the shortest, these two clusters are too large and should not be merged. The Point 6 and Point 7 are merged instead.

where ℓ is the softmax cross entropy loss. However, y_i is not available in the unsupervised setting, and it is challenging to find another objective function that can learn a robust embedding function ϕ.

7.3.2 The Bottom-up Clustering Framework

Without the manual annotation, it is important to design a supervision signal that can be used to train CNN models. To achieve this goal, we aim to exploit the similarity and diversity properties from the training data as the supervision information. As shown in Fig. 7.2 (a), the framework mainly contains two components: (i) A network trained with a repelled loss to let the cluster centers repelled by each other. (ii) A clustering procedure in the feature embeddings space to merge existing clusters. The clustering and network updating is done iteratively.

Network with Repelled Loss.

Since we do not have ground truth labels, we assign each image to a different cluster initially, *i.e.,* $\{\hat{y}_i = i \mid 1 \leq i \leq N\}$.[1] In this way, the network learns to recognize each training sample instead of the identities, and the diversity over each training sample is maximized. We then gradually incorporate similarity within identities by grouping similar images into clusters. The cluster ID is used as the training label, and the network is trained to minimize total intra-cluster variance and maximize the inter-cluster variance.

We define the probability that image x belongs to the c-th cluster as,

$$p(c|x,\boldsymbol{V}) = \frac{exp(\boldsymbol{V}_c^{\mathrm{T}}\boldsymbol{v}/\tau)}{\sum_{j=1}^{C} exp(\boldsymbol{V}_j^{\mathrm{T}}\boldsymbol{v}/\tau)}, \tag{7.2}$$

where $\boldsymbol{v} = \frac{\phi(\boldsymbol{\theta};x)}{||\phi(\boldsymbol{\theta};x)||}$, $\boldsymbol{V} \in \mathbb{R}^{C \times n_\phi}$ is a lookup table that stores the feature of each cluster, $\boldsymbol{V}_j$ is the j-th colum of $\boldsymbol{V}$, and C is the number of clusters at the current stage. At the first training stage, $C = |\boldsymbol{X}| = N$. At the following stages, our approach will merge similar images into one class, and C will gradually decrease. τ is a temperature parameter [30] that controls the softness of probability distribution over classes. Following [117], we set $\tau = 0.1$. In the forward operation, we compute cosine similarities between data x_i and all the other data by $\boldsymbol{V}^T \cdot \boldsymbol{v}_i$. During backward, we update the $\hat{y}_i$-th column of the table $\boldsymbol{V}$ by $\boldsymbol{V}_{\hat{y}_i} \leftarrow \frac{1}{2}(\boldsymbol{V}_{\hat{y}_i} + \boldsymbol{v}_i)$. Finally, we minimize the repelled loss, which is formulated as,

$$\mathcal{L} = -\log(p(\hat{y}_i|x_i,\boldsymbol{V})). \tag{7.3}$$

During the optimization, $\boldsymbol{V}_j$ will contain the information of all images within the j-th cluster. It can be considered as a kind of "centroid" of this cluster. We do not directly calculate the centroid feature in each training stage due to the high time complexity. The lookup table V can avoid exhaustive computation of extracting features from all data at each training step. The proposed objective has two advantages. First, it can maximize the cosine distance between each image feature $\boldsymbol{v}_i$ and each centroid features $\boldsymbol{V}_{j \neq \hat{y}_i}$. Second, it can minimize the cosine distance between each image feature $\boldsymbol{v}_i$ and the corresponding centroid feature $\boldsymbol{V}_{j=\hat{y}_i}$. With these two advantages, our approach can trade off the similarity and diversity over the whole training set.

[1] $\hat{y}_i$ is the cluster index for x_i and is dynamically changed.

Cluster Merging.

After the first training stage, the training samples are prone to be away from each other in the learned feature space. However, images of the same identity are usually visually similar and should be close, which we call *similarity*. To exploit the similarity, we apply the hierarchical clustering on the CNN features to merge the images from bottom to up. In the start, each image is treated as a cluster. Then pairs of clusters are merged into one by measuring their similarity. In order to decide which clusters should be merged, we consider the minimum distance criterion to calculate the dissimilarity value $D(A,B)$ between cluster A and cluster B.

The **minimum distance criterion** takes the shortest distance between images in two clusters as dissimilarity. This criterion only considers the shortest distance: if two images in the cluster look really alike, the clusters tend to be merged, no matter how dissimilar other images look. The advantage is that images of the same identity under the same camera are visually alike and tend to be merged into one cluster under this criterion, which guarantees the accuracy of merged images. It is formulated as:

$$D_{distance}(A,B) = \min_{x_a \in A, x_b \in B} d(x_a, x_b), \tag{7.4}$$

where $d(x_a, x_b)$ is defined as the Euclidean distance between the feature embeddings of two images, *i.e.*, $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$. Specifically, $d(x_a, x_b) = \|\boldsymbol{v}_a - \boldsymbol{v}_b\|$.

As shown in Fig. 7.2 (b)-(d), at each merging step, we aim to reduce m clusters. We define $m = N \times mp$, where $mp \in (0,1)$ denotes the speed of cluster merging. Each time, the clusters with the shortest distance are merged. The number of clusters is initialed as $C = N$, *i.e.*, the number of training samples. After t times of cluster merging, the number of clusters is dynamically decreased to $C = N - t \times m$.

There are other criteria methods to measure the dissimilarity. (1) The **maximum distance criterion** takes the maximum distance between elements of each cluster as the dissimilarity. However, images of the same identity under different cameras may have totally different visual appearance. This strategy fails to merge images from different cameras. (2) The **centroid distance criterion** takes the distance between mean features of elements in each cluster as the dissimilarity. In the re-ID task, images come from different cameras, which have different illumination, pose, and viewpoint. The mean operation omits the diversity among images within one cluster, therefore, it overlooks the useful camera information. In experiments, we demonstrate the minimum criterion is the best, and will discuss later in Section 7.4.3.

Algorithm 3: The Bottom-Up Clustering (BUC) Framework

```
Require: Unlabeled data $X = \{x_1, x_2, ...x_N\}$;
         Merge percent $mp \in (0,1)$;
         CNN model $\phi(\cdot;\boldsymbol{\theta}_0)$.
Ensure: Best CNN model $\phi(\cdot;\boldsymbol{\theta}^*)$.
Initialize: Cluster label $Y = \{\hat{y}_i | \ 1 \leq i \leq N\}$
            Number of cluster $C = N$
            Number of merging image $m = \lceil mp * C \rceil$
while $C > m$ do
    Train CNN model $\phi(x;\boldsymbol{\theta})$ with $X$ and $Y$
    Clustering with $m$:
        $C \leftarrow C - m$
        Update $Y$ with the new cluster labels
    Initialize the lookup table $\boldsymbol{V}$ with new dimensions
    Re-train the CNN model with parameters $\boldsymbol{\theta}$
    Evaluate on the validation set $\rightarrow$ performance $P$
    if $P > P^*$ then
        $P^* = P_t$
        Best model $= \phi(x;\boldsymbol{\theta})$
    end if
end while
```

Dynamic Network Updating.

The framework iteratively trains the network and merges the learned image features clusters. The clustering results are then fed to the network for further updating. The whole updating process is described in Algorithm 3. The number of clusters is initialized as the number of training images. After each cluster merging, the labels of the training images are re-assigned as the new cluster ID. The memory layer of the optimizer is randomly re-initialized to avoid getting stuck in local optima. We constantly train the network until we observe a performance drop on the validation set.

7.3.3 Diversity Regularization

With the clusters being merged, the number of classes is decreasing, and the number of images in the clusters is increasing. Although we do not know the exact number of images in each identity, we can assume that the images are evenly distributed to the identities, and different identities should be scattered in different clusters, which we call *diversity*. This implies that one cluster should not contain much more images compared to other clusters.

To avoid one cluster being redundant and boost the small clusters to merge together, we incorporate a diversity regularization term into the distance criterion.

$$D_{diversity}(A,B) = |A| + |B|, \tag{7.5}$$

where $|A|$ denotes the number of samples belonging to the cluster A. Then, the final dissimilarity is calculated as:

$$D(A,B) = D_{distance}(A,B) + \lambda D_{diversity}(A,B), \tag{7.6}$$

where λ is a parameter that balances the impact of distance and regularization. The reason for adding a diversity regularization term is that, there exist some visually similar identities wearing almost the same clothes. Without the regularization term, the algorithm might merge these similar but different identities into one tremendous cluster by mistake. We tend to merge small clusters, unless the distance $d(x_a, x_b)$ is small enough. This procedure is illustrated in Fig. 7.2 (d).

7.4 Experimental Results

7.4.1 Experimental Settings

Evaluation Protocols. For the image-based re-ID datasets Market-1501 and DukeMTMC-reID, we take all the training images without ID labels to train the framework. For the video-based datasets MARS and DukeMTMC-VideoReID, each training tracklet is regarded as an individual sample in the model training. Note, our method does not utilize any annotation information (*e.g.* ID labels or other annotated datasets) for model initialization or training.

Implementation Details. We adopt ResNet-50 as the CNN backbone to conduct all the experiments. We initialize it by the ImageNet [41] pre-trained model with the last classification layer removed. For all the experiments if not specified, we set the number of training epochs in the first stage to be 20, the batch size to be 16, the dropout rate to be 0.5, mp to be 0.05 and λ in Eq. (7.6) to be 0.005. We use stochastic gradient descent with a momentum of 0.9 to optimize the model. The learning rate is initialized to 0.1 and changed to 0.01 after 15 epochs. For video-based datasets, we take the average feature of all frames within a tracklet to be the tracklet feature for cluster merging and final evaluation. On Market-1501 and DukeMTMC-reID, it takes

Table 7.1 Comparison with the state-of-the-art methods on the Market-1501 dataset. The column "Labels" lists the labels utilized by the method. "Transfer" denotes the information from another re-ID dataset with full annotations. "OneEx" denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us. "No DR" denotes results without diversity regularization term.

Methods	Venue	Labels	rank-1	rank-5	rank-10	mAP
BOW [142]	ICCV15	**None**	35.8	52.4	60.3	14.8
OIM* [117]	CVPR18	**None**	38.0	58.0	66.3	14.0
UMDL [80]	CVPR16	Transfer	34.5	52.6	59.6	12.4
PUL [28]	TOMM18	Transfer	44.7	59.1	65.6	20.1
EUG* [114]	CVPR18	OneEx	49.8	66.4	72.7	22.5
SPGAN [11]	CVPR18	Transfer	58.1	76.0	82.7	26.7
TJ-AIDL [105]	CVPR18	Transfer	58.2	-	-	26.5
BUC No DR	AAAI19	**None**	62.9	77.1	82.7	33.8
BUC	AAAI19	**None**	**66.2**	**79.6**	**84.5**	**38.3**

Table 7.2 Comparison with the state-of-the-art methods on the DukeMTMC-reID dataset. The column "Labels" lists the labels utilized by the method. "Transfer" denotes the information from another re-ID dataset with full annotations. "OneEx" denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us. "No DR" denotes results without diversity regularization term.

Methods	Venue	Labels	rank-1	rank-5	rank-10	mAP
BOW [142]	ICCV15	**None**	17.1	28.8	34.9	8.3
OIM* [117]	CVPR18	**None**	24.5	38.8	46.0	11.3
UMDL [80]	CVPR16	Transfer	18.5	31.4	37.6	7.3
PUL [28]	TOMM18	Transfer	30.4	46.4	50.7	16.4
EUG* [114]	CVPR18	OneEx	45.2	59.2	63.4	24.5
SPGAN [11]	CVPR18	Transfer	46.9	62.6	68.5	26.4
TJ-AIDL [105]	CVPR18	Transfer	44.3	-	-	23.0
BUC (No DR)	AAAI19	**None**	41.3	55.8	62.5	22.5
BUC	AAAI19	**None**	**47.4**	**62.6**	**68.4**	**27.5**

about 4 hours to finish the training procedure with a GTX 1080TI GPU. On Mars and DukeMTMC-VideoReID, it takes about 5 hours.

7.4.2 Comparison with the State of the Art

Image-based Person Re-identification.

The comparisons with the state-of-the-art algorithms on image-based datasets are shown in Table 7.1 and Table 7.2. Note that the performances in [80] are reproduced by [28] and we borrow the numbers to our table. On Market-1501, we obtain the best

Table 7.3 Comparison with the state-of-the-art methods on MARS. The column "Labels" lists the labels utilized by the method. "OneEx" denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us. "No DR" denotes results without diversity regularization term.

Methods	Venue	Labels	rank-1	rank-5	rank-10	mAP
OIM* [117]	CVPR18	**None**	33.7	48.1	54.8	13.5
DGM+IDE [129]	ICCV17	OneEx	36.8	54.0	-	16.8
Stepwise [66]	ICCV17	OneEx	41.2	55.5	-	19.6
RACE [127]	ECCV18	OneEx	43.2	57.1	62.1	24.5
DAL [8]	BMVC18	Camera	49.3	65.9	72.2	23.0
EUG [114]	CVPR18	OneEx	**62.6**	74.9	-	**42.4**
BUC (No DR)	AAAI19	**None**	55.5	71.3	76.1	31.9
BUC	AAAI19	**None**	61.1	**75.1**	**80.0**	38.0

performance among the compared methods with **rank-1 = 66.2%, mAP = 38.3%**. Compared to the state-of-the-art method OIM [117] in the fully unsupervised setting, we achieve 28.2 points (absolute) and 24.3 points improvement in rank-1 accuracy and mAP, respectively. Similarly, our method achieves 22.9 points (absolute) and 16.2 points improvement in rank-1 and mAP on DukeMTMC-reID. The significant improvement is mainly due to the further cluster merging that exploits similarity from the instances for supervision.

We also compare our method to the state-of-the-art transfer learning methods in Table 7.1. Although these methods utilize external images and human annotations, our method with *zero* annotation still surpasses them by a large margin. On Market-1501, our method outperforms the state-of-the-art transfer learning method [105] by 8.0 points and 11.8 points in rank-1 accuracy and mAP, respectively.

Video-based Person Re-identification.

Table 7.3 and Table 7.4 shows the comparisons with the state-of-the-art algorithms on video-based datasets. On MARS, we obtain **rank-1 = 61.1%, mAP = 38.0%**. We beat the fully-unsupervised method OIM [117] by a large margins with 27.4 points in rank-1 accuracy and 24.5 points for mAP. On DukeMTMC-VideoReID, our results achieve 18.1 points and 18.1 points improvement on rank-1 accuracy and mAP, respectively.

In Table 7.3 and Table 7.4, we also compare our method to the state-of-the-art methods [66, 127, 129] in the video-based one-example setting. These methods initialize their models by annotating each person with a labeled video tracklet. As discussed in [114], these approaches are the one-example methods, hence, are not actually

Table 7.4 Comparison with the state-of-the-art methods on DukeMTMC-VideoReID. The column “Labels” lists the labels utilized by the method. “OneEx” denotes the one-example annotation, in which each person in the dataset is annotated with one labeled example. $*$ denotes that the results are reproduced by us. “No DR” denotes results without diversity regularization term.

Methods	Venue	Labels	rank-1	rank-5	rank-10	mAP
OIM* [117]	CVPR18	**None**	51.1	70.5	76.2	43.8
DGM+IDE [129]	ICCV17	OneEx	42.3	57.9	69.3	33.6
Stepwise [66]	ICCV17	OneEx	56.2	70.3	79.2	46.7
EUG [114]	CVPR18	OneEx	**72.7**	**84.1**	-	**63.2**
BUC (No DR)	AAAI19	**None**	60.7	76.8	80.6	50.8
BUC	AAAI19	**None**	69.2	81.1	**85.8**	61.9

unsupervised. Their methods rely on some very useful annotations on the dataset, *i.e.*, how many identities exist in the dataset and what they look like (from a tracklet for each person). Without any annotation, our method still beats most of these methods with one-example annotation, which indicates that our method is more effective in exploiting the unlabeled data.

7.4.3 Ablation Studies

The Impact of Diversity Regularization.

The performance of with and without the diversity regularization item is shown in Table 7.1 and Table 7.3, respectively. The diversity regularization provides a large performance improvement on all the four datasets. Specifically, on Market-1501 and DukeMTMC-reID, the diversity regularization item improves the rank-1 accuracy by 3.3 points and 6.1 points, respectively. We suspect that without the diversity regularization, two similar identities may be easily merged into one cluster by mistake. With the diversity regularization term, we tend to merge small clusters first.

The diversity regularization parameter λ in Eq. (7.6) balances the cluster size and cluster distance. We evaluate different values for the parameter λ in Fig. 7.3 (a). As λ increases from 0 to 0.005, the rank-1 accuracy on Market-1501 increases from 62.9% to 66.2%. If we set λ to be greater than 0.05, the too large diversity regularization term would begin to introduce a negative effect.

The Impact of Cluster Merging Criterion.

As shown in Table 7.5, the results of three cluster merging criteria are listed. We get the best result with the rank-1 = 66.2% when using the minimum distance criterion.

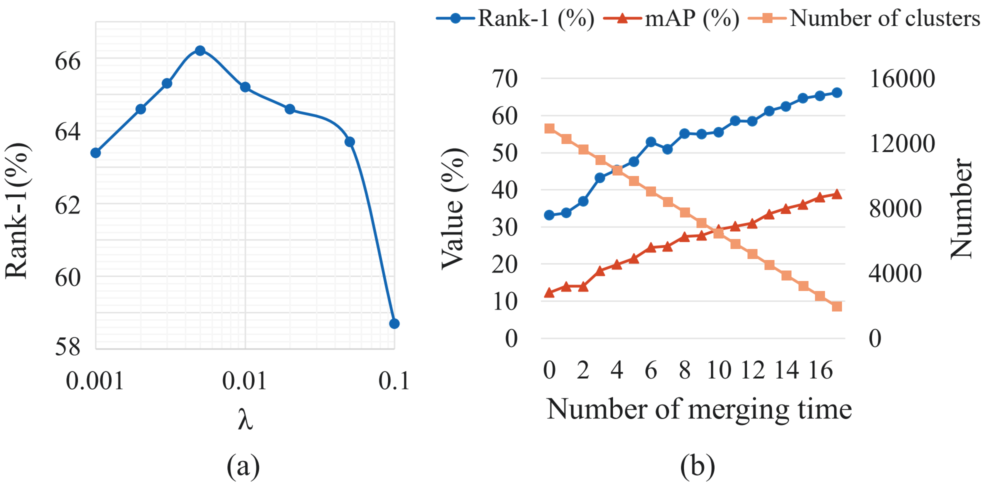


Fig. 7.3 (a) Performance curve with different values of the diversity regularization parameter λ on Market-1501. (b) The rank-1 accuracy, mAP, and the number of clusters on Market-1501 after each cluster merging step.

Table 7.5 The comparison of different merging criteria on Market-1501.

Criterion	rank-1	rank-5	rank-10	rank-20	mAP
Maximum	62.5	76.8	82.6	87.1	35.0
Centroid	65.8	79.2	83.6	88.4	37.9
Minimum	**66.2**	**79.6**	**84.5**	**88.5**	**38.3**

When using the centroid distance criterion, we observe a slightly lower performance with the rank-1 = 65.8%. When using the maximum distance criterion, we observe a rank-1 accuracy of 62.5%. We assume that images of the same identity from different cameras suffer from large visual appearance difference. Using this criterion may fail to merge clusters including images captured from different cameras.

7.4.4 Algorithm Analysis

Analysis over Cluster Merging.

We show the performance of re-ID and the number of remaining clusters on Market-1501 in Fig. 7.3 (b). As the number of the remaining clusters gradually decreases, the rank-1 accuracy and the mAP accuracy are both increasing. After 16 times of merging, the rank-1 accuracy increases from 33.2% to 66.2%, and the mAP accuracy increases from 12.3% to 38.3%. The number of clusters is decreased from 12,936 to 2,023, while

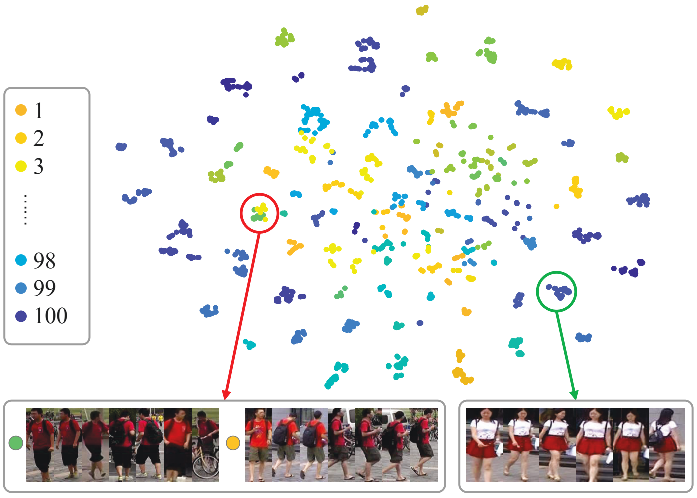


Fig. 7.4 T-SNE visualization of the learned feature embeddings on a part of the Market-1501 training set (100 identities, 1,700 images). Points of the same color represent images of the same identity. We show the detailed images of a positive example (the green circle) and a negative example (the red circle). The points in the green circle are of the same identity. In the red circle, the green and yellow points gathered together, indicating that our algorithm merges them into one cluster by mistake. However, the samples looks very similar and are hard to be discriminated between each other.

the ground truth number of identities is 751. We observe that both the improvement of the performance and the reduction of the clusters are continuous and gradual. It indicates that our method gradually learns from the diversified images to generate a more discriminative feature representation.

Qualitative Analysis.

To further understand the discriminative ability of our unsupervised learned feature, we utilize t-SNE [72] to visualize the feature embeddings of the merged clusters by plotting them to the 2-dimension map. As illustrated in Fig. 7.4, the images of the same identity usually gather together, which represents the learned similarity within identities. Besides, most identities are distinguishable from each other, which represents the diversity among the identities. More qualitative results over iterations can be found in the supplementary material.

Table 7.6 The top-1 accuracy on CIFAR-10.

Methods	top-1
S-CNN [21]	72.7
NID [115]	80.8
Roto-Scat + SVM [89]	82.3
DCGAN [81]	82.8
Ours	**85.2**

7.4.5 Compare to Unsupervised Feature Learning

To compare with the unsupervised feature learning methods, we also conduct image classification experiments on CIFAR-10 [40] to make a fair comparision with them. CIFAR-10 contains 60,000 images of 10 different classes. Following [115], we take ResNet18 as the backbone model and extract the last pooling layer's features. The nearest neighbor classifier is adopted to assess the learned feature, which reflects the quality of the representation. As shown in Table 7.6, we achieve 85.2% top-1 accuracy on CIFAR-10, showing a 4.4% accuracy gain over [115]. This improvement demonstrates the superiority of the cluster merging and network updating strategies.

7.5 Conclusions

In this chapter, we propose a bottom-up clustering approach (BUC) to tackle the unsupervised re-ID task. It jointly optimizes a CNN model and the relationship among the individual samples. Specifically, the network training starts by treating each individual image as an identity. Then, bottom-up clustering is applied to the feature embedding extracted from the network to reduce the number of classes. During the whole process, the network gradually exploits similarity from diverse unlabeled images. In experiments, BUC achieves higher performance than the state-of-the-art methods in both image-based and video-based re-ID datasets.

Chapter 8

Conclusion and Future Directions

In this thesis, we investigate the deep learning approaches to person re-identification. We contribute in three settings, including supervised learning, one-example learning and unsupervised learning.

For the first setting, we introduce a Bayesian based query expansion method and an attribute based method. We show that query expansion with reliable retrieved images could improve the re-ID baselines. We also investigate the person attributes and find that the identity labels and the attribute labels are complementary that leveraging these two type of labels could improve the performance of both re-ID and attribute prediction.

In one-example learning, we introduce two methods. The first method utilizes the labeled samples to initialize a network and progressively exploit unlabeled samples with the reliable pseudo label for further training. The second method improves the first method that further utilize the unselected samples for training.

In unsupervised learning, the previous methods mainly focus on unsupervised domain adaptation. However, we propose a bottom-up clustering method that iteratively exploits the similarity and maximize the diversity among the unlabeled images to learn the discriminative embeddings.

In the future, we will continue focusing on weakly supervised and unsupervised re-ID. Since cross-camera annotation is expensive, and it is exhausting to annotate images/videos for the real-world data, we aim to achieve the satisfied re-ID result with only a few or no labeled data. It is valuable to investigate the correlation of images captured by different cameras. Different from identity labels that need manual labeling, the camera ID is easy to obtain. We can learn from the difference between different cameras to solve the cross-camera re-ID problem under the unsupervised setting. Besides, without any annotation, we can exploit supervision from the image by self-supervise. For example, we could apply transformation such as rotation or scaling to the original image and learn from the difference and similarity of the images. The part-based person re-ID is also a promising direction. There have been many works that investigate part-based re-ID in a supervised setting and achieves impressive performances. We could also use the local patches as the auxiliary information of the global image for unsupervised learning.

References

[1] Abdulnabi, A. H., Wang, G., Lu, J., and Jia, K. (2015). Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959.

[2] Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916.

[3] Bak, S. and Carr, P. (2017). One-shot metric learning for person re-identification. In *CVPR*.

[4] Bautista, M. A., Sanakoyeu, A., Tikhoncheva, E., and Ommer, B. (2016). Cliquecnn: Deep unsupervised exemplar learning. In *NIPS*.

[5] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *ICML*.

[6] Chen, D., Yuan, Z., Chen, B., and Zheng, N. (2016a). Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268 – 1277.

[7] Chen, S.-Z., Guo, C.-C., and Lai, J.-H. (2016b). Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367.

[8] Chen, Y., Zhu, X., and Gong, S. (2018). Deep association learning for unsupervised video person re-identification. In *BMVC*.

[9] Chum, O., Mikulik, A., Perdoch, M., and Matas, J. (2011). Total recall ii: Query expansion revisited. In *CVPR*, pages 889 – 896.

[10] Chum, O., Philbin, J., Sivic, J., Isard, M., and Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, pages 1–8.

[11] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*.

[12] Deng, Y., Luo, P., Loy, C. C., and Tang, X. (2014). Pedestrian attribute recognition at far distance. In *ACMMM*, pages 789–792.

[13] Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

[14] Dong, X., Meng, D., Ma, F., and Yang, Y. (2017). A dual-network progressive approach to weakly supervised object detection. In *ACMMM*.

[15] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*.

[16] Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications TOMCCAP*, 14(4).

[17] Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360 – 2367.

[18] Figueira, D., Bazzani, L., Minh, H. Q., Cristani, M., Bernardino, A., and Murino, V. (2013). Semi-supervised multi-feature learning for person re-identification. In *AVSS*.

[19] Franco, A. and Oliveira, L. (2017). Convolutional covariance features: Conception, integration and performance in person re-identification. *Pattern Recognition*, 61:593–609.

[20] Garcia, J., Martinel, N., Micheloni, C., and Gardel, A. (2015). Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV*, pages 1305–1313.

[21] Ghaderi, A. and Athitsos, V. (2016). Selective unsupervised feature learning with convolutional neural network (s-cnn). In *ICPR*.

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*, pages 2672–2680.

[23] Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275.

[24] Han, D. and Kim, J. (2015). Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*.

[25] Hariharan, B., Malik, J., and Ramanan, D. (2012). Discriminative decorrelation for clustering and classification. In *ECCV*.

[26] He, J., Zhang, C., Zhao, N., and Tong, H. (2005). Boosting web image search by co-ranking. In *Acoustics, Speech, and Signal Processing*, volume 2, page 409.

[27] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778.

[28] Hehe, F., Liang, Z., Chenggang, Y., and Yi, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

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

[30] Hinton, G., Vinyals, O., and Dean, J. (2014). Distilling the knowledge in a neural network. In *NIPS-Workshop*.

[31] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *CVPR*.

[32] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

[33] Jain, V. and Varma, M. (2011). Learning to re-rank: query-dependent image re-ranking using click data. In *WWW*, pages 277–286.

[34] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACMMM*, pages 675–678. ACM.

[35] Jiang, L., Meng, D., Yu, S.-I., Lan, Z., Shan, S., and Hauptmann, A. (2014). Self-paced learning with diversity. In *NIPS*.

[36] Jose, C. and Fleuret, F. (2016). Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, pages 875–890.

[37] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *NIPS*, pages 3581–3589.

[38] Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.

[39] Kodirov, E., Xiang, T., and Gong, S. (2015). Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*.

[40] Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

[41] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*.

[42] Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *NIPS*.

[43] Layne, R., Hospedales, T. M., and Gong, S. (2014). Re-id: Hunting attributes in the wild. In *BMVC*.

[44] Layne, R., Hospedales, T. M., Gong, S., and Mary, Q. (2012). Person re-identification by attributes. In *BMVC*, volume 2, page 8.

[45] Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML-workshop*, volume 3, page 2.

[46] Lee, K. S., Croft, W. B., and Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *AVSS*, pages 235–242.

[47] Leng, Q., Hu, R., Liang, C., Wang, Y., and Chen, J. (2015). Person re-identification with content and context re-ranking. *Multimedia Tools and Applications*, 74(17):6989–7014.

[48] Li, D., Chen, X., Zhang, Z., and Huang, K. (2017). Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*.

[49] Li, J., Ma, A. J., and Yuen, P. C. (2018a). Semi-supervised region metric learning for person re-identification. *International Journal of Computer Vision*, pages 1–20.

[50] Li, W., Wu, Y., Mukunoki, M., and Minoh, M. (2012). Common-near-neighbor analysis for person re-identification. In *ICIP*, pages 1621–1624.

[51] Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152 – 159.

[52] Li, Z. and Tang, J. (2015). Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Transactions on Image Processing*.

[53] Li, Z., Tang, J., and Mei, T. (2018b). Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*.

[54] Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197 – 2206.

[55] Lin, Y., Dong, X., Zheng, L., Yan, Y., and Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 2.

[56] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., and Yang, Y. (2017). Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*.

[57] Lisanti, G., Masi, I., Bagdanov, A. D., and Del Bimbo, A. (2015). Person re-identification by iterative re-weighted sparse ranking. *IEEE T-PAMI*.

[58] Liu, C., Change Loy, C., Gong, S., and Wang, G. (2013). Pop: Person re-identification post-rank optimisation. In *ICCV*, pages 441 – 448.

[59] Liu, H., Feng, J., Qi, M., Jiang, J., and Yan, S. (2017a). End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506.

[60] Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., and Feng, J. (2017b). Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1.

[61] Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*.

[62] Liu, X., Song, M., Zhao, Q., Tao, D., Chen, C., and Bu, J. (2012). Attribute-restricted latent topic model for person re-identification. *Pattern recognition*, 45(12):4204–4213.

[63] Liu, Y. and Mei, T. (2011). Optimizing visual search reranking via pairwise learning. *IEEE Transactions on Multimedia*, 13(2):280–291.

[64] Liu, Y., Mei, T., Hua, X.-S., Tang, J., Wu, X., and Li, S. (2008). Learning to video search rerank via pseudo preference feedback. In *ICME*, pages 297–300.

[65] Liu, Y., Yan, J., and Ouyang, W. (2017c). Quality aware network for set to set recognition. In *CVPR*, pages 5790–5799.

[66] Liu, Z., Wang, D., and Lu, H. (2017d). Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*.

[67] Lv, J., Chen, W., Li, Q., and Yang, C. (2018). Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7948–7956.

[68] Ma, A. J. and Li, P. (2014a). Query based adaptive re-ranking for person re-identification. In *ACCV*, pages 397–412.

[69] Ma, A. J. and Li, P. (2014b). Semi-supervised ranking for re-identification with few labeled image pairs. In *ACCV*.

[70] Ma, F., Meng, D., Xie, Q., Li, Z., and Dong, X. (2017). Self-paced co-training. In *ICML*, pages 2275–2284.

[71] Ma, L., Yang, X., and Tao, D. (2014). Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670.

[72] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*.

[73] Martinel, N., Dunnhofer, M., Foresti, G. L., and Micheloni, C. (2017). Person re-identification via unsupervised transfer of learned visual representations. In *ICDSC*, pages 151–156.

[74] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363 – 1372.

[75] Matsukawa, T. and Suzuki, E. (2016). Person re-identification using cnn features learned from combination of attributes. In *ICPR*, pages 2428–2433.

[76] McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *CVPR*.

[77] Mei, T., Rui, Y., Li, S., and Tian, Q. (2014). Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(3):38.

[78] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS-W*.

[79] Peng, P., Tian, Y., Xiang, T., Wang, Y., and Huang, T. (2016a). Joint learning of semantic and latent attributes. In *ECCV*, pages 336–353.

[80] Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., and Tian, Y. (2016b). Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*.

[81] Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

[82] Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015). Semi-supervised learning with ladder networks. In *NIPS*, pages 3546–3554.

[83] Ren, L., Lu, J., Feng, J., and Zhou, J. (2017). Multi-modal uniform deep learning for rgb-d person re-identification. *Pattern Recognition*, 72:446–457.

[84] Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV-workshop*, pages 17–35.

[85] Rohini, U. and Varma, V. (2007). A novel approach for re-ranking of search results using collaborative filtering. In *ICCTA*, pages 491–496.

[86] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[87] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *NIPS*, pages 2234–2242.

[88] Schumann, A. and Stiefelhagen, R. (2017). Person re-identification by deep learning attribute-complementary information. In *CVPR-Workshop*, pages 1435–1443.

[89] Singh, A. and Kingsbury, N. (2017). Dual-tree wavelet scattering network with parametric log transformation for object classification. In *ICASSP*.

[90] Singh, S., Gupta, A., and Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *ECCV*.

[91] Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L. S., and Gao, W. (2018a). Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1167–1181.

[92] Su, C., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *ECCV*, pages 475–491.

[93] Su, C., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2018b). Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 75:77–89.

[94] Su, C., Zhang, S., Yang, F., Zhang, G., Tian, Q., Gao, W., and Davis, L. S. (2017). Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition*, 66:4–15.

[95] Sudowe, P., Spitzer, H., and Leibe, B. (2015). Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV*, pages 87–95.

[96] Sun, Y., Zheng, L., Deng, W., and Wang, S. (2017). Svdnet for pedestrian retrieval. In *ICCV*, pages 3820–3828.

[97] Tang, H. and Liu, H. (2016). A novel feature matching strategy for large scale image retrieval. In *IJCAI*.

[98] Tian, Q., Sebe, N., Lew, M. S., Loupias, E., and Huang, T. S. (2001). Image retrieval using wavelet-based salient points. *Journal of Electronic Imaging*, 10(4):835–850.

[99] Ustinova, E., Ganin, Y., and Lempitsky, V. (2017). Multi-region bilinear convolutional neural networks for person re-identification. In *AVSS*, pages 1–6.

[100] Varior, R. R., Haloi, M., and Wang, G. (2016a). Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808.

[101] Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016b). A siamese long short-term memory architecture for human re-identification. In *ECCV*, pages 135–153.

[102] Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *AVSS*, pages 141 – 147.

[103] Wang, H., Gong, S., and Xiang, T. (2014). Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*.

[104] Wang, H., Gong, S., Zhu, X., and Xiang, T. (2016a). Human-in-the-loop person re-identification. In *ECCV*, pages 405–422.

[105] Wang, J., Zhu, X., Gong, S., and Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*.

[106] Wang, T., Gong, S., Zhu, X., and Wang, S. (2016b). Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514.

[107] Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018a). Person transfer gan to bridge domain gap for person re-identification. In *CVPR*.

[108] Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018b). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88.

[109] Wengert, C., Douze, M., and Jégou, H. (2011). Bag-of-colors for improved image search. In *ACMMM*, pages 1437–1440.

[110] Wu, L., Shen, C., and Hengel, A. v. d. (2016). Deep recurrent convolutional networks for video-based person re-identification: an end-to-end approach. *arXiv preprint arXiv:1606.01609*.

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

[112] Wu, L., Wang, Y., Gao, J., and Li, X. (2018a). Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition*, 73:275–288.

[113] Wu, Y., Lin, Y., Dong, X., Yan, Y., Bian, W., and Yang, Y. (2019). Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing*.

[114] Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018b). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

[115] Wu, Z., Xiong, Y., Stella, X. Y., and Lin, D. (2018c). Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.

[116] Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249 – 1258.

[117] Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *CVPR*.

[118] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR*.

[119] Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *AVSS*, pages 211 –215.

[120] Xu, J., Zhao, R., Zhu, F., Wang, H., and Ouyang, W. (2018). Attention-aware compositional network for person re-identification. In *CVPR*.

[121] Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., and Zhou, P. (2017). Jointly attentive spatial-temporal pooling networks for video-based person re-identification. *ICCV*.

[122] Yamamoto, T., Nakamura, S., and Tanaka, K. (2007). Rerank-by-example: Efficient browsing of web search results. In *DEXA*, pages 801–810.

[123] Yan, C., Luo, M., Liu, W., and Zheng, Q. (2018). Robust dictionary learning with graph regularization for unsupervised person re-identification. *Multimedia Tools and Applications*.

[124] Yan, R. and Hauprmann, A. (2007). Query expansion using probabilistic local feedback with application to multimedia retrieval. In *CIKM*, pages 361–370.

[125] Yan, R., Hauptmann, A., and Jin, R. (2003). Multimedia search with pseudo-relevance feedback. In *ICIVR*, pages 238–247.

[126] Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., and Pan, Y. (2012). A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):723–742.

[127] Ye, M., Lan, X., and Yuen, P. C. (2018). Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*.

[128] Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., and Hu, R. (2016). Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566.

[129] Ye, M., Ma, A. J., Zheng, L., Li, J., and Yuen, P. C. (2017). Dynamic label graph matching for unsupervised video re-identification. *ICCV*.

[130] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *ICPR*.

[131] Yin, Z., Zheng, W.-S., Wu, A., Yu, H.-X., Wan, H., Guo, X., Huang, F., and Lai, J. (2018). Adversarial attribute-image person re-identification. In *IJCAI*, pages 1100–1106.

[132] Yu, H.-X., Wu, A., and Zheng, W.-S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*.

[133] Zhang, L., Xiang, T., and Gong, S. (2016). Learning a discriminative null space for person re-identification. In *ICCV*, pages 1239 – 1248.

[134] Zhang, S., Yang, M., Cour, T., Yu, K., and Metaxas, D. N. (2012). Query specific fusion for image retrieval. In *ECCV*, pages 660–673.

[135] Zhang, S., Yang, M., Cour, T., Yu, K., and Metaxas, D. N. (2015). Query specific rank fusion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):803–815.

[136] Zhang, W., Hu, S., and Liu, K. (2017). Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*.

[137] Zhao, L., Li, X., Zhuang, Y., and Wang, J. (2017). Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239 – 3248.

[138] Zhao, R., Ouyang, W., and Wang, X. (2013a). Person re-identification by salience matching. In *ICCV*, pages 2528–2535.

[139] Zhao, R., Ouyang, W., and Wang, X. (2013b). Unsupervised salience learning for person re-identification. In *CVPR*.

[140] Zhao, R., Ouyang, W., and Wang, X. (2014). Learning mid-level filters for person re-identification. In *CVPR*, pages 144 – 151.

[141] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016a). Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884.

[142] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *CVPR*, pages 1116 – 1124.

[143] Zheng, L., Wang, S., Liu, Z., and Tian, Q. (2014). Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR*, pages 1947 – 1954.

[144] Zheng, L., Yang, Y., and Hauptmann, A. G. (2016b). Person re-identification: past, present and future. *arXiv preprint arXiv:1610.02984*.

[145] Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649 – 656.

[146] Zheng, Z., Zheng, L., and Yang, Y. (2017a). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):13.

[147] Zheng, Z., Zheng, L., and Yang, Y. (2017b). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3774 – 3782.

[148] Zhong, Z., Lei, M., Cao, D., Fan, J., and Li, S. (2017a). Class-specific object proposals re-ranking for object detection in automatic driving. *Neurocomputing*, 242:187–194.

[149] Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017b). Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661.

[150] Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y. (2018). Camera style adaptation for person re-identification. In *CVPR*.

[151] Zhou, S., Wang, J., Meng, D., Xin, X., Li, Y., Gong, Y., and Zheng, N. (2018). Deep self-paced learning for person re-identification. *Pattern Recognition*, 76:739–751.

[152] Zhou, Z., Huang, Y., Wang, W., Wang, L., and Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785.

[153] Zhu, J., Liao, S., Lei, Z., and Li, S. Z. (2017a). Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229.

[154] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017b). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232.

[155] Zhu, X., Jing, X. Y., Yang, L., You, X., Chen, D., Gao, G., and Wang, Y. (2017c). Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1.

[156] Zhu, X., Wu, B., Huang, D., and Zheng, W.-S. (2018). Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 27(5):2286–2300.