Visual Analysis with Limited Supervision

Wenhe Liu

Supervisor: Professor Yi Yang

Centre for Artificial Intelligence
Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of
Doctor of Philosophy

November 2018

I would like to dedicate this thesis to my loving parents and my beloved grandparents.

Declaration

I, Wenhe Liu declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Wenhe Liu
November 2018

Acknowledgements

This thesis cannot be completed without important people in my study and life. And I would like to acknowledge my supervisor, Dr. Yi Yang, who led me into the realm of machine learning and visual analysis. His passion for exploring new ideas and rigorous attention to detail affected me profoundly. Dr. Xiaojun Chang is my associate supervisor. I am always impressed by the way he thinks about every research problem. His self-motivation, hard-working spirit and rigorous research attitude have set an excellent example for me. Dr. Alexander G. Hauptmann was my supervisor when I visited Carnegie Mellon University. I have appreciated his guidance immensely. I would like to thank many friends at University of Technology Sydney and Carnegie Mellon University. Yan Yan, Guoliang Kang have helped me so much when I was in the group. Their sincerity and kind heart impressed me to give warm care to every other person. I would also like to thank my friends at University of Adelaide. Dong Gong, a good friend who has taught me a lot of mathematical theory and skills and support me all the time. Lastly, I would like to thank my parents for their continuous support for my academic pursuit. I wish them health and happiness forever.

Abstract

Visual analysis is an attractive research topic in the field of computer vision. In the visual analysis, there are two critical directions, visual retrieval and visual classification. In recent years, visual retrieval has been investigated and developed in many real-world applications, for instance, in person re-identification. On the other hand, visual classification is also widely studied, such as in image classification. Typical visual analysis methods are supervised learning algorithms. In such algorithms, extensive labeled data is demanded for training supervised models in order to achieve acceptable performance. However, it is difficult to collect and generate annotated data in the real world due to the limited resources, such as human labor for annotation. Therefore, it is urgent to develop methods to complete the visual analysis mission with limited supervision.

In this thesis, we propose to address the visual analysis problem with limited supervision. Specifically, we treat limited supervision problem in three scenarios according to the amount of labeled data. In the first scenario, no labeled data are provided and only limited human labor for annotation is available; In the second scenario, scarce labeled data and abundant unlabeled data are accessible. In the third scenario, only few instances in the target dataset are labeled and there are multiple sources of labeled data from different domains.

In Chapter 2 and Chapter 3, we discuss the first scenario, when no labeled data are provided, and only limited human labor for annotation is available. We propose to solve the problem via active learning. Unlike conventional active learning, which usually starts with a set of labeled data as the reference, in this thesis, we adopt the active learning algorithm with no pre-given labeled data. We refer these algorithms as the Early Active Learning. In this thesis, first, we attempt to select the most contributive instances for annotation and later being utilized for training supervised models. We demonstrate that even by annotating a few selected instances, the proposed method can achieve comparable performance in the visual retrieval. Second, we further extend the instance based active learning to pair-based early active learning. Other than

select instances for annotation, the pair-based early active learning selects the most informative pairs for annotation, which is essential in the visual retrieval.

In Chapter 4, in the second scenario, we address the visual retrieval problem when there are scarce labeled data and abundant unlabeled data. In this thesis, we propose to utilize both the labeled and the unlabeled data in a semi-supervised attribute learning schema. The proposed method could jointly learn the latent attributes with appropriate dimensions and estimate the pairwise probability of the data simultaneously.

In Chapter 5 and Chapter 6, in the third scenario, we focus on visual classification with few or no labels, but there are pre-known labeled data from other domains. To improve the performance in the target domain, we adopt transfer learning algorithms to transfer helpful knowledge from the pre-known (source) domain with labeled data. First, in Chapter 5, the few-shot visual classification problem is considered. We have access to multiple source datasets with well-labeled data but can only access a limited set of labeled data in the target dataset. An Analogical Transfer Learning schema is proposed for this problem. It attempts to transfer the knowledge from the source domains to enhance the performance of the target domain models. In the algorithm, an analogy-revision schema is designed to select only the helpful source instances to enhance the target domain models. Second, in Chapter 6, we challenge a more difficult problem when there is no labeled data in the target domain in the visual retrieval problem. A Domain-aware Unsupervised Cross-dataset Transfer Learning algorithm is proposed to address this problem. The importance of universal and domain-unique appearances are valued simultaneously and jointly contribute to the representation learning. It manages to leverage the common and domain-unique representations across datasets in the unsupervised visual retrieval.

Contents

List of Figures

List of Tables

Publications

From this work:

- Chapter 2: **Wenhe Liu**,Xiaojun Chang, Ling Chen, Yi Yang: Early Active Learning with Pairwise Constraint for Person Re-identification. *ECML/PKDD*, 2017
- Chapter 3: **Wenhe Liu**,Xiaojun Chang, Yi Yang:Pair-based Early Active Learning for Person Re-identification. (submitted 2018)
- Chapter 4: **Wenhe Liu**, Xiaojun Chang, Ling Chen, Yi Yang: Semi-Supervised Bayesian Attribute Learning for Person Re-Identification. *AAAI*, 2018.
- Chapter 5: **Wenhe Liu**, Xiaojun Chang , Yan Yan , Yi Yang , Alexander G. Hauptmann Few-Shot Text and Image Classification via Analogical Transfer Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Volume 9 Issue 6, October 2018.
- Chapter 6: **Wenhe Liu**,Xiaojun Chang, Yi Yang:Domain-aware Unsupervised Cross-dataset Transfer Learning for Person Re-identification. (submitted 2018)

Referred:

- **Wenhe Liu**, Dong Gong, Mingkui Tan, Javen Qinfeng Shi, Yi Yang and Alexander G. Hauptmann: Learning Distilled Graph for Large-scale Social Network Data Clustering. *Transactions on Knowledge and Data Engineering (TKDE).* (submitted 2017)
- **Wenhe Liu**, Chenqiang Gao, Xiaojun Chang, Qun Wu: Unified discriminating feature analysis for visual category recognition. *J. Visual Communication and Image Representation*,2016.

Chapter 1

Introduction

1.1 Background

1.1.1 Video Analysis

Visual analysis is an attractive research topic in the field of computer vision. In recent years, visual analysis has been investigated and developed in many real-world applications. There are two impotent topics in the field of visual analysis, which are visual retrieval and visual classification.

Visual retrieval gains much attention surveillance computer vision, for instance, in person re-identification (Re-ID) [138]. In model Re-ID methods, such as in [48–50, 79, 138], Re-ID can be formed as an image *retrieval* task. Given a *probe* image of a person from one camera view, the difficulty is to identify images of the same person from a *gallery* of images taken by other non-overlapping camera views. Meanwhile, in recent researches [94, 102, 112, 119] visual classification is also studied widely on, e.g. image classification [109].

Supervised visual analysis methods can achieve promising results if there are sufficient labeled training data. In such algorithms, extensive labeled data is demanded when training supervised models in order to achieve acceptable performance. However, it is difficult to collect and generate annotated data in the real world due to the limitation of resources, such as human labor for annotation. Unfortunately, the human labor for labeling training data is sometimes inadequate in the real world. For instance, for visual retrieval, annotation of data becomes extremely severe in the Re-ID scenario, since annotation for pairwise labels is difficult to achieve. Re-ID data requires all pairs of images to be labeled. It is a tough task even for humans to compare and sort the images from a potentially massive number of imposters [49, 89].

1.1.2 Learning with Limited Supervision

It is essential to design effective algorithmS to complete the visual analysis mission with limited supervision. We consider three scenarios according to the amount of labeled data. In the first scenario, no labeled data are provided and only limited human labor for annotation is available; In the second scenario, scarce labeled data and abundant unlabeled data are accessible. In the third scenario, only few instances in the target dataset are labeled and there are multiple sources of labeled data from different domains.

The scenario is usually in the early stage of the experiments when there are no labeled data and only limited human annotators. In this scenario, we challenge the visual analysis problem when no labeled data are provided and only a limited human labor for annotation is available. We propose to solve the problem via active learning. These kinds of active learning algorithms are referred to as *early active learning* or *early stage experimental design* [85]. It attempts to select the most contributive instances for annotation and later being utilized for training supervised models. Compared to conventional active learning, early active learning claims it is developed for the early stage of experiments when there is no labeled data. We demonstrate that even annotate a few selected instances, the proposed method can achieve comparable performance. We develop an instance based early active learning and a pair-based early active learning for visual retrieval.

The second scenario is usually in the middle stage of the experiments when there are a number of labeled data (usually very few) and abundant unlabeled data, which is referred to as semi-supervised learning. The main challenge in this scenario is to utilize both the labeled and unlabeled data jointly to improve learning performance. To solve the visual retrieval problem, we proposed a Bayesian framework combine an Indian buffet process (IBP) [27] prior in an infinite latent factor model that enables adaptively learning attributes [6]. Additionally, we also define a probability for relation learning which utilizes both the labeled and unlabeled data.

The third scenario is usually in the late stage of the experiments when there are already multiple source domains with sufficient pre-known instances, and there is a new coming (target) domain with only a few instances known. The main difficulty of few-shot learning is how to optimize the target model when there come new classes of data, and only a few labeled training instances are provided for each class. Given sufficient pre-known labeled data from related domains, such a problem can be addressed by transfer learning (TL) [88]. Transfer learning benefits the target task by transferring helpful prior knowledge from some source domains. With the prior knowledge from the source domains, the performance of the learning task in the target domain could be improved even with few samples [16, 88]. Moreover, while most existing works rely on the abundance of labeled exemplars, we consider a more difficult unsupervised scenario, where no labeled exemplar is provided. One solution for unsupervised visual analysis that attracts much attention in the recent researches is cross-dataset transfer learning. It utilizes knowledge from multiple source datasets from different domains to enhance the unsupervised learning performance on the target domain. In this thesis, we propose a novel domain-aware representation learning algorithm for unsupervised cross-dataset transfer learning. The proposed algorithm not only learns a common appearances across-datasets but also captures the domain-unique appearances on the target dataset via minimization of the overlapped signal supports across different domains.

1.2 Related Works

1.2.1 Active Learning

To save labor costs, it is essential to design an effective algorithm that can select a subset of samples that are the most representative and/or informative for training. Active learning is widely studied to solve this kind of sample selection problem. As discussed in [85], active learning methods can be divided into two categories. The first category of algorithms select the most informative samples for labeling when there are already some labeled samples. They include uncertainty sampling methods [4, 40, 58, 111] query by committee methods [23, 100]. Most of these active learning methods prefer to select uncertainty data, or data that is difficult to analyze. They thus require a certain number of labeled samples to evaluate the uncertainty of the unlabeled data or sampling bias [85] will result. It is therefore recommended that such methods are only applied in the mid-stage of experiments when there are sufficient labeled data. For the purpose of distinguishing between the two categories, we refer to the first category of active learning methods as *traditional active learning.* The second category of active learning methods is considered for application in the early stage of experiments, when there are limited resources for labeling data. In this case, there are no labeled samples, thus labeling a small number of representative data is desirable for training reliable supervised models. In the category of early active learning, there are clustering-based methods [81, 86] and transductive experimental design methods [130]. These kinds of active learning algorithms are referred to as *early active learning* or *early stage experimental design* [85].

1.2.2 Semi-supervised Attribute Learning in Re-id

Attribute learning. One straightforward solution in representation learning for visual analysis is to learn the most representative features directly from images. Factor-based representation, such as dictionary learning [47, 50], has shown promising performance. Additionally, in recent years, some approaches have begun to incorporate deep learning [120] with even better results. In some recently works [66, 103], it proposed manually annotating attributes for use in a deep learning framework. In [98, 104], deep learning models were designed to be trained on separate datasets with attribute labels, then fine-tuned on target datasets without attribute labels. However, pre-training such algorithms is still limited by the number and type of attributes in the datasets that have been manually annotated.

Supervised and semi-supervised learning in person Re-ID. Various traditional and state-of-the-art machine learning strategies are investigated to solve the Re-ID problem, such as distance metric learning [62, 63, 80, 108], deep learning[1, 120], learning to rank[68, 78, 79] and dictionary learning [60, 92]. Among all the existing methods, most of them are supervised methods. However, these methods are not scalable as they require a large amount of labeled training data. Supervised methods for Re-ID needs sufficient images of people to be matched across each pair of camera views. The annotation task is tough even for the human as it requires identifying the same person in different camera views from a huge number of imposters. Besides, sometimes people do not reappear in other camera views. As a result, the scalability of the supervised methods is limited for large-scale practical Re-ID applications [49].

Semi-supervised Re-ID methods also have been proposed in previous Re-ID works [49, 79, 113, 135]. Typical semi-supervised and unsupervised methods are offered for a single dataset with less or no label information and report a much weaker performance than supervised methods. The reason is that with insufficient labels of matching pair information, they are not efficient in learning appearance features when there is dramatically variance of data. Thus they can fail to recognize a person under severe appearance changes across camera views [89].Some semi-supervised Re-ID methods have also been proposed [73, 79]. Commonly, the training models in semi-supervised methods rely on both labeled and unlabeled data. Hence, they produce acceptable performance compared to supervised methods without an abundance of labeled data [79].

1.2.3 Transfer Learning and Few-shot Learning

Transfer learning. Transfer learning aim to improve the learning task in the target domain (with few instances) by transferring helpful prior knowledge from some source domains. Multi-source Transfer Learning focus on building an ensemble model from multiple source domains that suit for the machine learning task on target domain [64, 117]. There are also multi-task learning [9] which attempts to learn both the source and the target tasks simultaneously, while transfer learning only aims at improving the performance of the target domain tasks. Based on the knowledge level they transfer, transfer learning can be divided into three types: feature transfer learning (FTL), instance transfer learning (ITL) and parameter/model transfer learning (PTL) [88]. There are also a few joint transfer learning algorithms which joint two or more of the three algorithms, such as feature and instance transfer [37]. Some of TL algorithms consider to combine transfer learning with other learning algorithms, such as kernel learning methods [75] and metric learning methods [125].

Feature transfer learning. In FTL, the algorithms are designed to learn a feature representation in the source domain and use it for the target domain. The feature representations are desired to minimize the discrepancy between domains such that it could enhance the learning performance of the target task. For instance, there is Kernelized Bayesian Transfer Learning finds a shared subspace of source and target domain by a kernel-based Bayesian dimensionality reduction model [28].

Instance transfer learning. The main idea of ITL algorithms is to reweigh or reallocate samples in the source domain based on the target domain data [88]. Some classical works are proposed in [39, 42, 131]. In many works, the importance of the source instance is stated as the similarity between the source and the target domain data. Probabilistic methods evaluate the distribution similarity in [5, 93, 105]. Graph-based methods evaluate similarity by its local weights or structure similarity [19, 20, 24, 26, 38, 82, 95, 99]. Generally, the mentioned algorithms rely on the estimation of relatedness between the source and the target domain data. However, the dependence of similarity analysis also brings obstacles to such algorithms. Considerably different distributions of the source and the target domain can frustrate the measure of similarity [99].

Parameter/Model transfer learning. In PTL, knowledge is transferred in terms of parameter or hyperparameters of common distributions [88]. For instance, there is Projective Model Transfer learning which adopts a regularization term to standard SVM by analyzing the angle between hyperplanes of the source and the target domain models [3]. Hypothesis Transfer Learning (HTL) aims to improve target hypothesis by transferring source hypothesis. It does not require the estimation of relatedness of instances between domains. Plenty works have been explored reliable performance in the field [56, 87, 109]. In [53], the authors present a multi-class hypothesis transfer learning algorithm. It updates both the target and the pre-trained sources hypothesis when a new class of target data observed. A recent work [117] presents an unsupervised hypothesis transfer learning. In the work, pseudo labels are first generated to source domain samples. Hypothesis learning is then conducted on a model transfer SVM. In [15], an active selection strategy is used to select semantic constraints rather than transferring hypotheses. Recently, a simple to complex learning schema is presented in [70] on transfer learning for action recognition. In [71], the authors discuss the parameter stability for multi-task learning with shared similar feature structures. In [2], the effectiveness of the transfer to the target domain and compatibility of the transfer model HTL algorithms are analyzed. Overall, the mentioned methods consider the source models as well-trained and unchangeable.

Self-paced learning. Self-paced learning [29, 44, 52] is a study paradigm that could adaptively learn and select subsets of instances by easiness that improve the performance of main learning task. In self-paced learning, several discrete regularization terms are present and used in varies of applications [107, 134] For continues regularization, a work in [123] develops a logistic function related regular term. The diversity learning of SPL is present in [43, 131], however in discrete form. In our hypothesis learning algorithm, we use a specialized continues regular term to analyze and control negative transfers.

Few-shot/one-shot learning. Few-shot/one-shot learning [94, 102, 112, 119] aims to solve the problem that instead of given one large dataset, there are only a few annotated samples (or only one annotated example) for each class in training data. There is gradient-based optimization models [94], metric learning models [102, 112] and deep embedding methods[106].

1.3 Contributions

For the three scenarios of limited supervision, we propose different algorithms to address the visual analysis problems. The main contribution of this thesis is stated as follows:

In Chapter 2 and Chapter 3, this thesis is the first to propose the early active learning (EAL) methods for visual retrieval. Specifically, we apply two different kinds of EAL algorithms, which are instance-based and pair-based. The instance-based early active learning methods is "Early Active Learning with Pairwise Constraint". In the algorithm, a pairwise constraint is introduced to the conventional EAL algorithm for person re-id. The pair-based early active algorithm is "Pair-based Early Active Learning". In the algorithm, pairs of instances are evaluated with diversity maximization criterion to enhance the pairwise diversity of selected pairs of samples.

In Chapter 4, this thesis proposes Bayesian framework unifies representation learning and Re-ID probability estimation and can simultaneously optimize both learning tasks. In the algorithm, the dictionary of attributes is adaptively determined using an efficient estimation method.

In Chapter 5 , this thesis proposed a novel analogical transfer learning (ATL) algorithm. Rather than transferring knowledge from the source hypothesis to learn the target hypothesis, ATL learns an analogical hypothesis from both source and target hypothesis. Moreover, ATL is able to revise the source hypotheses by select helpful source instances according to their contribution to the target hypothesis. As a result, the proposed algorithm efficiently controls the occurrence of the negative transfer on both instance and hypothesis level.

In Chapter 6, this thesis proposes a novel unsupervised cross-dataset learning algorithm with support discriminative regularization for person Re-ID . To our knowledge, it is the first attempt to leverage the common and domain-unique representations across datasets in the unsupervised Re-ID application.

Chapter 2

Early Active Learning with Pairwise Constraint

2.1 Background

As we mentioned in Introduction, the primary target of person re-identification (Re-ID) is to identify a person from camera shots across pairs of non-overlapping camera views, and research on this topic has attracted considerable attention in recent years [48–50, 79, 138]. In the field of computer vision, Re-ID can be formed as an image *retrieval* task. Given a *probe* image of a person from one camera view, the difficulty is to identify images of the same person from a *gallery* of images taken by other non-overlapping camera views. Despite the encouraging results reported in previous works, Re-ID remains a challenge in several respects. The accuracy of identification is often degrades as a result of the uncontrollable and/or unpredictable variation of appearance changes across camera views, such as body pose, view angle, occlusion and illumination conditions [47, 89, 120].

Supervised Re-ID methods can achieve promising results if there are sufficient labeled training data. Unfortunately, the human labor necessary for labeling training data is sometimes inadequate. This problem becomes extremely severe in the Re-ID scenario, since labeling for Re-ID is difficult to achieve. Unlike other recognition tasks which only requires each image to be labeled, Re-ID requires all pairs of images across camera views to be labeled. It is a tough task even for humans to identify the same person in different camera views among a potentially huge number of imposters [49, 89]. At the same time, pairwise labeled data is required for each pair of camera views in the camera network in Re-ID , thus the labeling cost will become prohibitively given the large numbers of cameras in today's world. For example, there might be more than over a hundred in one underground train station [89].

To save labor costs, it is essential to design an effective algorithm that can select a subset of samples that are the most representative and/or informative for training. Active learning is widely studied to solve this kind of sample selection problem. As discussed in [85], active learning methods can be divided into two categories. The first category of algorithms select the most informative samples for labeling when there are already some labeled samples. They include uncertainty sampling methods [4, 40, 58, 111] query by committee methods [23, 100]. Most of these active learning methods prefer to select uncertainty data, or data that is difficult to analyze. They thus require a certain number of labeled samples to evaluate the uncertainty of the unlabeled data or sampling bias [85] will result. It is therefore recommended that such methods are only applied in the mid-stage of experiments when there are sufficient labeled data. For the purpose of distinguishing between the two categories, we refer to the first category of active learning methods as *traditional active learning.* The second category of active learning methods is considered for application in the early stage of experiments, when there are limited resources for labeling data. In this case, there are no labeled samples, thus labeling a small number of representative data is desirable for training reliable supervised models. In the category of early active learning, there are clustering-based methods [81, 86] and transductive experimental design methods [130]. These kinds of active learning algorithms are referred to as *early active learning* or *early stage experimental design* [85]. We illustrate the procedures of and example of the traditional active learning algorithm, QUIRE [40], and our early active learning algorithm with pairwise constraint (abbreviated as EALPC) in Fig. 2.1.

In the rest of this chapter, we focus on the early active learning methods for person re-identification applications. As mentioned, labeling Re-ID data is extremely labor-consuming and time-consuming. It is therefore highly desirable to enhance the learning performance in Re-ID applications by early active learning. Unfortunately, early active learning methods currently merely consider analyzing representative samples with pairwise relationships. Therefore, directly applying them for Re-ID may be not appropriate.

To overcome the limitations described above, we propose a novel algorithm for person re-identification, Early Active Learning with Pairwise Constraint, abbreviated as EALPC. The main contributions of this chapter are as follows:

1. We propose a novel Early Active Learning with Pairwise Constraint algorithm for person re-identification. To the best of our knowledge, this is the first

method considers to consider both (a) applying early active learning for the Re-ID application, and (b) extending early active learning schema with pairwise constraint.

2. We introduce the $\ell_{2,1}$-norm to our objective function, which improves the robustness of our methods and suppresses the effects of outliers.

3. We propose an efficient algorithm to optimize the proposed problem. Our optimization algorithm also provides a closed form solution and guarantees to reach the global optimum in the convergence.

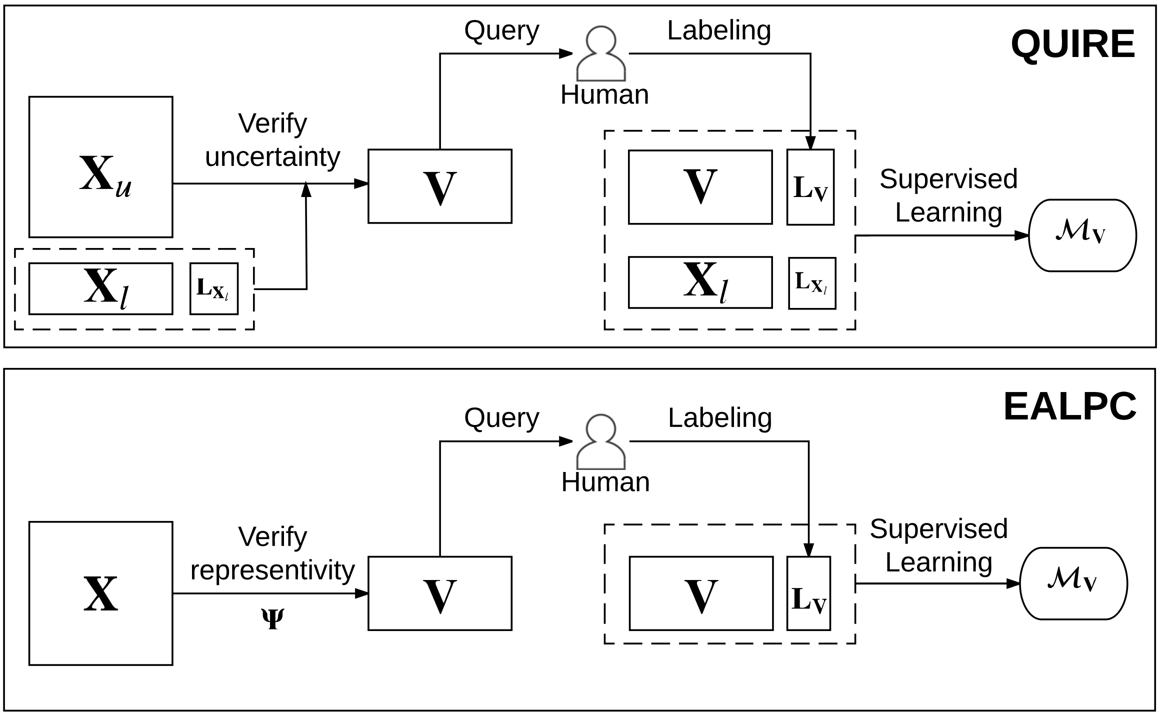


Figure 2.1 Procedures of QUIRE [40] (*upper*) and our Early Active Learning with Pairwise Constraint (EALPC) (*lower*). In QUIRE, pre-labeled samples $\mathbf{X}_l$ are used for the uncertainty evaluation on the unlabeled samples $\mathbf{X}_u$. Then, it selects a subset samples $\mathbf{V} \subset \mathbf{X}_u$ for labeling. At last, both $\mathbf{X}_u$ and $\mathbf{V}$ along with their labels are used for supervised learning. In EALPC, unlabeled data $\mathbf{X}$ is analyzed without pre-labeled data. Meanwhile, pairwise constraint $\mathbf{\Psi}$ is introduced to enhance the performance of early active learning for Re-ID . More details are in Section. 2.2.

2.2 The Proposed Method

In this section, we first revisit the early active learning algorithm and then propose our early active learning with pairwise constraint for Re-ID .

Notation

Let the superscript $^{\mathsf{T}}$ denote the transpose of a vector/matrix, $\mathbf{0}$ be a vector/matrix with all zeros, $\mathbf{I}$ be an identity matrix. Let $\mathrm{Tr}(\mathbf{A})$ be the trace of matrix $\mathbf{A}$. Let $\mathbf{a}_i$ and $\mathbf{a}^j$ be the i-th column vector and j-th row vector of matrix $\mathbf{A}$ respectively. Let $\langle \mathbf{A},\mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}\mathbf{B}^{\mathsf{T}})$ be the inner product of $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{v}\|_p$ be the ℓ_p-norm of a vector $\mathbf{v}$. Then, the Frobenius norm of an arbitrary matrix $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A},\mathbf{A}\rangle}$. The ℓ_2-norm of a vector $\mathbf{a}$ is denoted as $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}}$ and the $\ell_{2,1}$-norm of matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is denoted as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{n}\|\mathbf{a}^i\|_2$, where a_{ij} is the (i,j)-th element of $\mathbf{A}$ and $\mathbf{a}^i$ is the i-th row vector of $\mathbf{A}$. For analytical consistency, the $\ell_{2,0}$-norm of a matrix $\mathbf{A}$ is denoted as the number of the nonzero rows of $\mathbf{A}$. Let $\mathrm{vec}(\mathbf{A})$ be a column vector generated from the matrix $\mathbf{A}$ by concatenating all column vectors of $\mathbf{A}$. Let $\mathrm{diag}(\mathbf{v})$ be a diagonal matrix with diagonal elements equal to $\mathbf{v}$, $\langle \mathbf{A},\mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}\mathbf{B}^{\mathsf{T}})$ be the inner product of $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{v}\|_p$ be the ℓ_p-norm of a vector $\mathbf{v}$. For any convex function $f(\mathbf{A})$, let $\partial f(\mathbf{A})/\partial\mathbf{A}$ denote its subdifferential at $\mathbf{A}$.We denote $\mathcal{G}$ as a weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ constructed on $\mathcal{X}$. The (unnormalized) Laplacian matrix associated with $\mathcal{G}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a degree matrix with $\mathbf{D}(i,i) = \sum_j S(i,j)$. Let $\mathbf{a}\circ\mathbf{b}$ represent the Hadamard (element-wise) product between two vectors $\mathbf{a}$ and $\mathbf{b}$. Let $\mathbf{a}^{\odot 2} = \mathbf{a}\circ\mathbf{a}$ be the element-wise square of $\mathbf{a}$.

2.2.1 Early Active Learning

We first revisit the early active learning algorithm [85]. Given a set of unlabeled samples $\mathbf{X} \in \mathbb{R}^{d\times n}$, the task of active learning is to select a subset of $m < n$ most representative samples $\mathbf{V} \in \mathbb{R}^{d\times m}$. Then, the selected samples are queried labeling for supervised learning. The labeled subset of data is expected to maximize the potential performance of the supervised learning in the early stage of experiment, when the available resource for labeling data is limited, i.e. only a small number of data can be labeled for supervised learning. Generally, we can define the optimization problem of early active learning as follows:

$$\min_{\mathbf{V},\mathbf{A}} \mathbf{R}(\mathbf{X},\mathbf{V},\mathbf{A}) + \alpha\Omega(\mathbf{A}), \;\; s.t. \;\mathbf{V} \subset \mathbf{X}, \; |\mathbf{V}| = m. \tag{2.1}$$

where $\mathbf{V}$ is a subset of $\mathbf{X}$, $\mathbf{A}$ is a transformation matrix. In Eq. (2.1), the first term $\mathbf{R}(\cdot)$ is the reconstruction loss, the second term $\Omega(\cdot)$ is the regularization term and $\alpha > 0$ is a leverage parameter.The major purpose of early active learning is to select a subset $\mathbf{V} \subset \mathbf{X}$ with size $m < n$ that can best represent the whole data $\mathbf{X}$ through the linear transformation matrix $\mathbf{A}$. The selected samples are therefore considered to be the most representative.

In [130], an early active learning via a Transduction Experimental Design algorithm (TED) is proposed with the aim of finding the subset $\mathbf{V} \subset \mathbf{X}$ and a project matrix $\mathbf{A}$ that minimizes the least squared reconstruction error:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \sum_{i=1}^{n} (\|\mathbf{x}_i - \mathbf{V}\mathbf{a}_i\|_2^2 + \alpha \|\mathbf{a}_i\|_2^2) \\ &\qquad s.t. \;\; \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \; \mathbf{V} \subset \mathbf{X}, \; |\mathbf{V}| = m. \end{aligned} \tag{2.2}$$

where $\mathbf{V}\mathbf{a}_i$ is the representation item of $\mathbf{x}_i$. However, Eq. (2.2) is an NP-hard problem to solve, thus an approximate solution by a sequential optimization problem is proposed in [130].

2.2.2 Early Active Learning with Pairwise Constraint

In this chapter, we focus on early active learning in the person Re-ID problem. As mentioned previously, person Re-ID is formed as an image *retrieval* task which aims to re-identify the same person across non-overlapping camera views given a probe image of the person. The analysis of pairwise relationships of images in different camera views is therefore required. For this purpose, we introduce a pairwise constraint to early active learning:

$$\Psi_{\mathbf{V}}(\mathbf{A}) = \sum_{i,j=1}^{n} \|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|_2^2 S_{\mathbf{V}}(i,j), \tag{2.3}$$

where $\mathbf{V}\mathbf{a}_i$ is the representation item of $\mathbf{x}_i$ and $S_{\mathbf{V}}(i,j)$ is the (i,j)-th element of similarity matrix $\mathbf{S}$. It is the similarity between the i-th and the j-th representations. In this chapter we define $S_{\mathbf{V}}(i,j)$ as a Gaussian similarity:

$$S_{\mathbf{V}}(i,j) = \begin{cases} \exp(-\frac{\|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|^2}{\sigma^2}), & \textit{if } \mathbf{V}\mathbf{a}_i \in \mathcal{N}_k(\mathbf{V}\mathbf{a}_j) \textit{ and } \mathbf{V}\mathbf{a}_j \in \mathcal{N}_k(\mathbf{V}\mathbf{a}_i) \\ 0 & , \textit{otherwise}, \end{cases} \tag{2.4}$$

where $\mathcal{N}_k(\mathbf{x})$ denotes the set of k-nearest neighbors of $\mathbf{x}$. We can then reformulate the pairwise constraint in Eq. (2.3) by inducing a Laplacian matrix:

$$\Psi_{\mathbf{V}}(\mathbf{A}) = \sum_{i,j=1}^{n} \|\mathbf{V}\mathbf{a}_i - \mathbf{V}\mathbf{a}_j\|_2^2 S_{\mathbf{V}}(i,j) = \mathrm{Tr}((\mathbf{V}\mathbf{A})\mathbf{L}_{\mathbf{V}}(\mathbf{V}\mathbf{A})^T), \tag{2.5}$$

where $\mathbf{L}_{\mathbf{V}} = \mathbf{D} - \mathbf{S}_{\mathbf{V}}$ is the Laplacian matrix and $\mathbf{D}$ is the degree matrix with each element $\mathbf{D}_{ii} = \sum_j S_{\mathbf{V}}(i,j)$. As discussed in [49], minimizing the pairwise constraint will force the similar representations to be close to each other. Following the assumption that visually similar images of a person have a high probability of sharing the similar representation features in Re-ID [49], this will make early active learning schema more suitable for Re-ID applications.

After introducing the pairwise constraint, the early active learning for person re-identification can be formulated as:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \mathbf{R}(\mathbf{X},\mathbf{V},\mathbf{A}) + \alpha\Omega(\mathbf{A}) + \beta\Psi_{\mathbf{V}}(\mathbf{A}) \\ &\qquad s.t. \quad \mathbf{A} = [\mathbf{a}_1,\cdots,\mathbf{a}_n] \in \mathbb{R}^{m\times n}, \mathbf{V} \subset \mathbf{X},\ |\mathbf{V}| = m. \end{aligned} \tag{2.6}$$

where $\alpha > 0$ and $\beta > 0$ are leverage parameters of regularization terms. After substituting Eq. (2.2) and Eq. (2.5) into Eq. (2.6) we obtain:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{A}} \sum_{i=1}^{n} (\|\mathbf{x}_i - \mathbf{V}\mathbf{a}_i\|_2^2 + \alpha\|\mathbf{a}_i\|_2^2) + \beta\mathrm{Tr}((\mathbf{V}\mathbf{A})\mathbf{L}_{\mathbf{V}}(\mathbf{V}\mathbf{A})^T) \\ &\qquad s.t. \quad \mathbf{A} = [\mathbf{a}_1,\cdots,\mathbf{a}_n] \in \mathbb{R}^{m\times n}, \mathbf{V} \subset \mathbf{X},\ |\mathbf{V}| = m. \end{aligned} \tag{2.7}$$

Finding the optimal subset $\mathbf{V} \subset \mathbf{X}$ in Eq. (2.7) is NP-hard. Inspired by [85], we relax the problem to the following problem by introducing the $\ell_{2,0}$-norm for structure sparsity:

$$\begin{aligned} &\min_{\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \alpha\|\mathbf{A}\|_{2,0} + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T) \\ &\qquad s.t. \quad \mathbf{A} = [\mathbf{a}_1,\cdots,\mathbf{a}_n] \in \mathbb{R}^{n\times n},\ \|\mathbf{A}\|_{2,0} = m. \end{aligned} \tag{2.8}$$

However, the $\ell_{2,0}$-norm makes Eq. (2.8) a non-convex problem. At the same time, the least squared loss used in Eq. (2.8) is sensitive to the outliers [85], which makes the algorithm not robust.

We note that in previous researches [84, 85, 128], the $\ell_{2,1}$-norm is used instead of the $\ell_{2,0}$-norm. It is shown in [85] that the $\ell_{2,1}$-norm is the minimum convex hull of the $\ell_{2,0}$-norm when row-sparsity is required. In other words, minimization of $\|\mathbf{A}\|_{2,1}$ will achieve the same result as $\|\mathbf{A}\|_{2,0}$ when $\mathbf{A}$ is row-sparse. As analyzed in [85, 139], the $\ell_{2,1}$-norm can suppress the effect of outlying samples. We therefore reformulate Eq. (2.8) as a relaxed convex optimization problem:

$$\min_{\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \tag{2.9}$$

In Eq. (2.9), we adopt the $\ell_{2,1}$-norm instead of both the least square reconstruction loss term and the $\ell_{2,0}$-norm structure sparsity term for robustness and suppression of outliers. By inducing the matrix formulation, Eq. (2.9) is rewritten as follows:

$$\min_{\mathbf{A}} \|(\mathbf{X} - \mathbf{X}\mathbf{A})^T\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \tag{2.10}$$

After obtaining the optimal solution of $\mathbf{A}$, the importances of samples can be ranked by sorting the absolute row-sum values of $\mathbf{A}$ in the decreasing order. A subset of the representative samples then can be selected corresponding to the top m largest values and query labeling.

2.2.3 Kernelization

The proposed algorithm can be extended to the kernel version for non-linear high dimensional space. We define $\Phi : \mathbb{R}^d \rightarrow \mathcal{H}$ as a mapping from the Euclidian space to a Reproducing Kernel Hilbert Space (RKHS) as $\mathcal{H}$. It can be induced by a kernel function $\mathcal{K}(\mathbf{x},\mathbf{y}) = \Phi(\mathbf{x})^T\Phi(\mathbf{y})$. Then we can project $\mathbf{X}$ to RKHS space as $\Phi(\mathbf{X}) = [\Phi(\mathbf{x}_1), \cdots, \Phi(\mathbf{x}_n)]$. The proposed problem thus becomes:

$$\min_{\mathbf{A}} \|(\Phi(\mathbf{X}) - \Phi(\mathbf{X})\mathbf{A})^T\|_{2,1} + \alpha\|\mathbf{A}\|_{2,1} + \beta \mathrm{Tr}((\Phi(\mathbf{X})\mathbf{A})\mathbf{L}_{\mathbf{X}}(\Phi(\mathbf{X})\mathbf{A})^T). \tag{2.11}$$

We denote our Early Active Learning with Pairwise Constraint algorithm in Eq. (2.10) as EALPC and the kenerlized version of our algorithm in Eq. (2.11) as EALPC_K.

2.2.4 Optimization

We provide an efficient algorithm for optimizing the proposed objective function. Taking the derivative w.r.t. $\mathbf{A}$ in Eq. (2.10) and setting it to zero, we obtain [1]:

$$\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{P} - \mathbf{X}^T\mathbf{X}\mathbf{P} + \alpha\mathbf{Q}\mathbf{A} + \beta\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{L}_{\mathbf{X}} = \mathbf{0}, \tag{2.12}$$

where $\mathbf{P}$ is a diagonal matrix and its i-th diagonal element is $p_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$. $\mathbf{Q}$ is a diagonal matrix and its i-th diagonal element is $q_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. Then by setting the derivative of Eq. (2.12) w.r.t. $\mathbf{a}_i$ to zero for each i, we obtain:

$$p_{ii}\mathbf{X}^T\mathbf{X}\mathbf{a}_i - p_{ii}\mathbf{X}^T\mathbf{x}_i + \alpha\mathbf{Q}\mathbf{a}_i + \beta\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{L}_i = \mathbf{0}, \tag{2.13}$$

where $\mathbf{L}_i$ is the i-th column vector of $\mathbf{L}_{\mathbf{X}}$. It is sample to verify that $\mathbf{A}\mathbf{L}_i = l_{ii}\mathbf{a}_i + \sum_{k\neq i} l_{ki}\mathbf{a}_k$, where l_{ii} and l_{ki} are the (i,i)-th and (k,i)-th element of $\mathbf{L}_{\mathbf{X}}$ respectively and $\mathbf{a}_k$ is the k-th column vector of $\mathbf{A}$. Therefore, the optimal solution $\mathbf{a}_i^*$ can be calculated by the closed form solution:

$$\mathbf{a}_i^* = (p_{ii}\mathbf{X}^T\mathbf{X} + \alpha\mathbf{Q} + \beta\mathbf{X}^T\mathbf{X}l_{ii})^{-1}(p_{ii}\mathbf{X}^T\mathbf{x}_i - \beta\mathbf{X}^T\mathbf{X}\sum_{k\neq i}\mathbf{a}_k l_{ki}). \tag{2.14}$$

In Eq. (2.12), $\mathbf{P}$ and $\mathbf{Q}$ are dependent on $\mathbf{A}$, thus they also need to be determined in each iteration. We propose an iterative algorithm to solve this problem. The detailed algorithm is described in Algorithm 1. In the next section, we will prove that Algorithm 1 converges to the global optimal solution of Eq. (2.10).

2.3 Convergence Analysis

We first introduce a lemma proposed in [84]:

Lemma 1. *For any arbitrary vector* $\mathbf{m}$ *and* $\mathbf{n}$ *there is*

$$\|\mathbf{m}\|_2 - \frac{\|\mathbf{m}\|_2^2}{2\|\mathbf{n}\|_2} \leq \|\mathbf{n}\|_2 - \frac{\|\mathbf{n}\|_2^2}{2\|\mathbf{n}\|_2}. \tag{2.15}$$

Next, in the following theorem we prove the convergence of our algorithm:

[1] In practice, when $\mathbf{x}_i - \mathbf{X}\mathbf{a}_i = 0$, p_{ii} can be regularized as $p_{ii} = \frac{1}{2\sqrt{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \eta}}$. Similarly when $\mathbf{a}_i = \mathbf{0}$, we set $q_{ii} = \frac{1}{2\sqrt{\|\mathbf{a}^i\|_2^2 + \eta}}$. η is a very small constant. It can be verified that when $\eta \to 0$ the problem with η reduces to the original problem in Eq. (2.12).

Algorithm 1: Algorithm for solving problem in Eq. (2.10)

```
  Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, parameters $\alpha$ and $\beta$.
1 Initialize $\mathbf{A} \in \mathbb{R}^{n\times n}$.
2 while not converge do
3     Compute the diagonal matrix $\mathbf{P}$, where the $i$-th diagonal element is
      $p_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$.
4     Compute the diagonal matrix $\mathbf{Q}$, where the $i$-th diagonal element is
      $q_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$.
5     Update $\mathbf{A}$ by each column $\mathbf{a}_i$ as in Eq. (2.14):
          $\mathbf{a}_i^* = (p_{ii}\mathbf{X}^T\mathbf{X} + \alpha\mathbf{Q} + \beta\mathbf{X}^T\mathbf{X}l_{ii})^{-1}(p_{ii}\mathbf{X}^T\mathbf{x}_i - \beta\mathbf{X}^T\mathbf{X}\sum_{k\neq i}\mathbf{a}_k l_{ki})$.
  Output: The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.
```

Theorem 1. *Algorithm 1 monotonically decreases the objective function value of Eq.* (2.10) *in each iteration.*

Proof. Suppose in an iteration the updated $\mathbf{A}$ is $\mathbf{A}^+$. According to Step 5 in Algorithm 1 we know that:

$$\mathbf{A}^+ = \arg\min_{\mathbf{F}} f(\mathbf{F}), \tag{2.16}$$

where we denote the function

$$f(\mathbf{F}) = \mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{F})\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{F})^T) + \alpha\mathrm{Tr}(\mathbf{F}\mathbf{Q}\mathbf{F}^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{F})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{F})^T).$$

Thus, in each iteration when updating $\mathbf{A}$ to $\mathbf{A}^+$ we have

$$\begin{aligned} &\mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{A}^+)\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{A}^+)^T) + \alpha\mathrm{Tr}((\mathbf{A}^+)\mathbf{Q}(\mathbf{A}^+)^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^+)\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^+)^T) \\ &\leq \mathrm{Tr}((\mathbf{X} - \mathbf{X}\mathbf{A})\mathbf{P}(\mathbf{X} - \mathbf{X}\mathbf{A})^T) + \alpha\mathrm{Tr}(\mathbf{A}\mathbf{Q}\mathbf{A}^T) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \end{aligned} \tag{2.17}$$

According to the definition of $\mathbf{P}$ and $\mathbf{Q}$, we thus obtain:

$$\begin{aligned} &\sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i^+\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \alpha\frac{\|\mathbf{a}^{i+}\|_2^2}{2\|\mathbf{a}^i\|_2}\right) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^+)\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^+)^T) \\ &\leq \sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \alpha\frac{\|\mathbf{a}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right) + \beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T). \end{aligned} \tag{2.18}$$

Meanwhile, according to Lemma 1, we can induce the following inequalities:

$$\sum_{i=1}^{n}\left(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i^{+}\|_2-\frac{\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i^{+}\|_2^2}{2\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2}\right)\leq\sum_{i=1}^{n}\left(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2-\frac{\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2}\right), \tag{2.19}$$

and

$$\sum_{i=1}^{n}\left(\|\mathbf{a}^{i+}\|_2-\frac{\|\mathbf{a}_i^{+}\|_2^2}{2\|\mathbf{a}_i\|_2}\right)\leq\sum_{i=1}^{n}\left(\|\mathbf{a}^{i}\|_2-\frac{\|\mathbf{a}^{i}\|_2^2}{2\|\mathbf{a}^{i}\|_2}\right). \tag{2.20}$$

After summing Eq. (2.19)and Eq. (2.20) in the both sides of Eq. (2.18), we conclude that:

$$\begin{aligned}&\sum_{i=1}^{n}(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i^{+}\|_2+\alpha\|\mathbf{a}^{i+}\|_2)+\beta\mathrm{Tr}((\mathbf{X}\mathbf{A}^{+})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A}^{+})^T)\\&\leq\sum_{i=1}^{n}\left(\|\mathbf{x}_i-\mathbf{X}\mathbf{a}_i\|_2+\alpha\|\mathbf{a}^{i}\|_2\right)+\beta\mathrm{Tr}((\mathbf{X}\mathbf{A})\mathbf{L}_{\mathbf{X}}(\mathbf{X}\mathbf{A})^T).\end{aligned} \tag{2.21}$$

The above inequality indicates that the objective function value of Eq. (2.10) monotonically decreases in Algorithm 1. □

Meanwhile, let $\partial f(\mathbf{A})/\partial\mathbf{A}=0$ is equal to solving Eq. (2.12), thus in convergence, $\mathbf{A}$ will satisfy Eq. (2.10). As Eq. (2.10) is a convex problem, $\mathbf{A}$ is the global optimum solution to our problem. Overall, Algorithm 1 will converge to the global optimum solution of Eq. (2.10).

2.4 Experimental Study

In the experiments, we compare our proposed EALPC algorithm with five state-of-the-art and classic active learning algorithms. After determining and labeling the most representative samples, we train the Re-ID models with these samples using five popular Re-ID algorithms. All experiments are operated on four widely referenced Re-ID benchmark datasets. We report the average performance of 10 trials of independent experiments on each dataset.

Datasets and Algorithms

1). Datasets. We analyze performance of active learning for Re-ID on four widely referred benchmark datasets for person re-identification. **VIPeR:** [30] The VIPeR dataset contains 1,264 images of 632 persons from two non-overlapping camera views. Two images are taken for each person, each from a different camera. Variations in viewpoint and illumination conditions occur frequently in VIPeR. **PRID:** [35] The PRID dataset contains images of 385 individuals from two distinct cameras. Camera B records 749 persons and Camera A records 385 persons, 200 of whom are same persons. **i-LID:** [139] The i-LID dataset records 119 individuals captured by three different cameras in an airport terminal. It contains 476 images with large occlusions caused by luggage and viewpoint changes. **CAVIAR:** [14] The CAVIAR dataset contains 72 individuals captured by two cameras in a shopping mall. The number of the images is 1,220, with 10 to 20 images for each individual. The size of the images in the CAVIAR dataset varies significantly from 39×17 to 141×72.

In the experiments, we use the recently proposed Local Maximal Occurrence (LOMO) features for person image representation [62]. As in [69, 89], all person images are scaled to 128×48 pixels. We then use the default setting in [62] to produce a 29,960 dimension feature for each image.

2). Active Learning Algorithms. We choose five active learning algorithms and compare them with our proposed algorithm. **Random:** As a baseline algorithm, we randomly select samples and query labeling. **K-means:** We use the K-means algorithm as another baseline algorithm as in [85]. In each experiment, samples are ranked by their distances from the K cluster centers in ascending order. **QUIRE:** [40] Active learning by Querying Informative and Representative Examples is an algorithm which queries the most informative and representative examples for labeling using the min-max margin-based approach. **TED:** [130] Active learning via Transduction Experimental Design is an algorithm that selects a subset of informative samples from a candidate dataset. It formulates a regularized linear regression problem which minimizes reconstruction error. **RRSS:** [85] Early active learning via Robust Representation and Structured Sparsity is a early active learning algorithm. It uses the $\ell_{2,1}$-norm to introduce structured sparsity for sample selection and robustness. However, RRSS does not consider the pairwise relations in Re-ID . We also introduce the kernelized RRSS denoted as **RRSS_K**. **EALPC:** Our proposed early active learning with pairwise constraint algorithm is denoted as EALPC. We also use a kernelized version of our algorithm denoted as **EALPC_K:**. For kernelization, we construct a Gaussian kernel for the candidate dataset, i.e. $\mathcal{K}(x_i, x_j) = \exp(-\alpha \|x_i - x_j\|^2)$. To seek the optimal parameters (if any), we apply a grid search in a region of $\{10^{-4}, 10^{-3}, \cdots, 1, \cdots, 10^3, 10^4\}$ with a five-fold cross validation strategy to determine the best parameters.

3). Re-identification Algorithms. Five state-of-the-art supervised Re-ID algorithms are chosen for the performance analysis of the proposed early active learning algorithms on person Re-ID . **NFST:** [132] Null Foley-Sammon Transform space learning is a Re-ID algorithm for learning a discriminative subspace where the training data points of each of the classes are collapsed to a single point. **KCCA:** [69] Kernel Canonical Correlation Analysis algorithm seeks a common subspace between the proposed images extracted from disjoint cameras and projects them into a new space. **XQDA:** [62] Cross-view Quadratic Discriminant Analysis learns a discriminant low dimensional subspace by cross-view quadratic discriminant analysis for metric learning. **kLFDA:** [122] Kernelized Local Fisher Discriminant Classifier is a closed form method that uses a kernelized method to handle large dimensional feature vectors while maximizing a Fischer optimization criterion. **MFA:** [126] Marginal Fisher Analysis method is introduced for dimensionality reduction by designing two graphs that characterize the intra-class compactness and interclass separability.

4). Settings. We report the average performance of 10 independent trials. In each trial, we divide each dataset into two equal-sized subsets as training and test sets, with no overlapping of person identities. Following the setting in [89], we divide the probe and gallery sets for Re-ID as follows: for datasets recording two camera views, e.g. VIPeR and PRID, images of one view are randomly selected for the probe sets, and images from the other view are chosen for the gallery sets. For a multi-view dataset, e.g. i-LID, images of one view are randomly selected as gallery sets and others are chosen as probe images. For the training set, we apply active learning methods to select a subset of training samples and query human labeling. The supervised Re-ID algorithms are then trained with the labeled samples. For evaluation measurement, we evaluate the performance of Re-ID by Cumulative Matching Characteristic (CMC) curve, which is the most commonly used performance measure for person Re-ID algorithms [47, 62, 68]. CMC calculates the probability that there exists a candidate image in the rank k gallery set that appears to match the prob image. In the experimental study, we also report the Rank One Matching Accuracy from CMC for simplicity.

Experimental Result Analysis

1). Performance of Re-id. We illustrate the performance of the active learning algorithms for Re-ID application in Table 2.1. In Table 2.1, each row corresponds to an active learning algorithm, and each column corresponds to a supervised Re-ID method. On each benchmark dataset, we select 20% of training samples via active learning algorithms and query labeling. The labeled subsets of samples are then adopted by supervised Re-ID algorithms for training models. We report the rank one matching accuracy in Table 2.1.

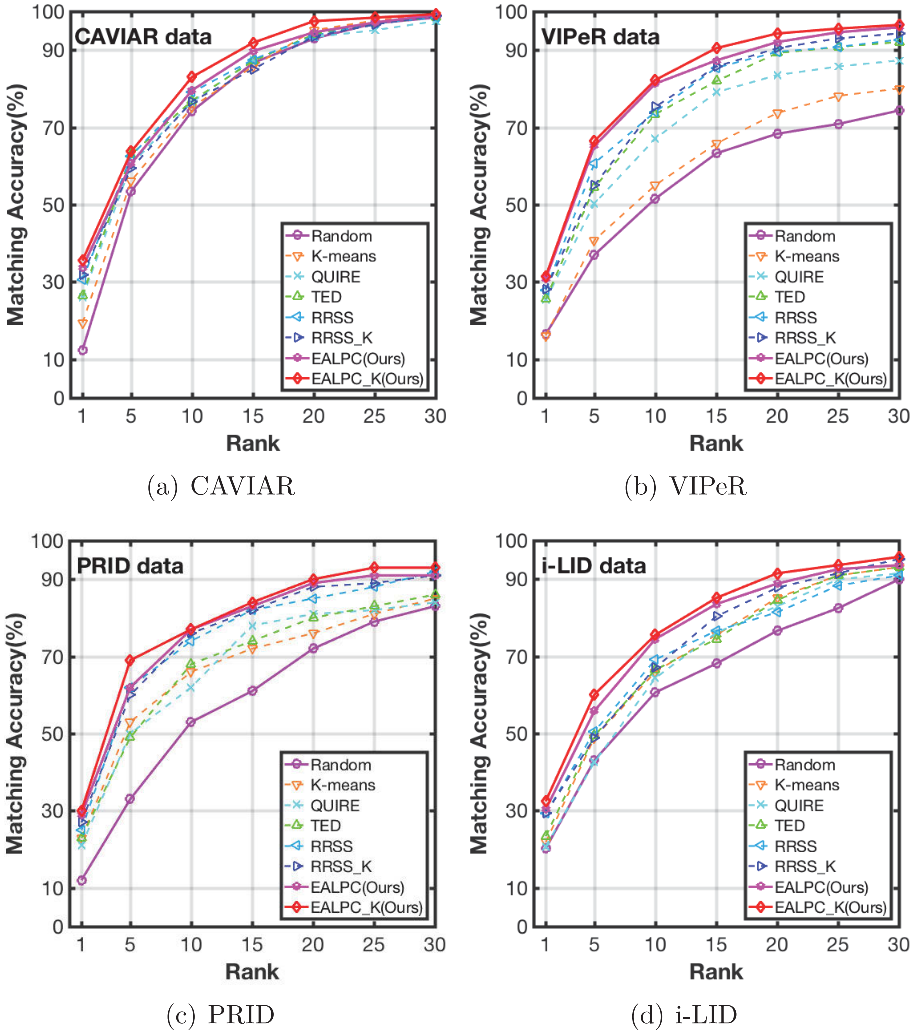
As shown in Table 2.1, we observe that: 1) All active learning algorithms perform better than Random selection. This indicates that active learning algorithms can select useful samples to improve the performance of Re-ID . 2) Our algorithms consistently outperform the other active learning algorithms. The table also confirms that our

Dataset	CAVIAR					VIPeR				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	23.65	23.47	21.38	27.55	25.87	26.65	23.01	27.23	22.78	23.64
K-means	26.90	25.99	22.05	27.74	27.40	27.59	26.16	27.59	23.15	24.39
TED	29.78	28.70	29,42	27.94	28.08	27.45	28.53	28.43	25.75	26.09
QUIRE	30.66	30.87	31.56	28.18	26.16	28.39	27.43	28.54	26.25	25.13
RRSS	31.87	30.69	33.57	30.95	29.01	31.56	28.54	30.71	27.34	28.04
RRSS_K	31.69	33.03	35.56	31.41	31.13	31.61	28.73	31.46	28.51	29.40
EALPC	34.12	33.57	37.45	33.09	31.16	32.61	29.45	31.82	28.54	29.56
EALPC_K	**35.00**	**35.20**	**38.75***	**33.29**	**31.91**	**33.66**	**30.44**	**34.29***	**29.18**	**30.03**

Dataset	PRID					iLIDS				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	24.49	25.47	24.00	23.50	20.00	25.96	23.40	25.00	23.35	25.00
K-means	26.16	27.54	27.01	24.70	21.20	27.02	23.94	27.00	25.57	25.20
TED	27.72	27.71	29.32	24.33	22.11	29.15	25.33	28.13	27.33	29.20
QUIRE	27.24	26.90	29.33	24.40	22.50	28.72	25.74	28.03	29.48	30.20
RRSS	29.21	28.44	30.00	25.09	23.97	28.11	27.66	30.82	30.08	30.55
RRSS_K	30.33	29.03	31.05	25.30	24.10	29.17	27.37	32.00	30.30	31.10
EALPC	32.22	30.63	31.03	25.90	25.60	29.26	27.66	32.34	30.43	31.60
EALPC_K	**32.70**	**31.50**	**33.40***	**26.06**	**25.70**	**31.19**	**28.72**	**34.00***	**31.60**	**32.47**

Table 2.1 Rank One Matching Accuracy(%) on four benchmarks. Percentage of selected instances for labeling is 20% of all samples. Each column is an active learning algorithm and each row is a Re-ID algorithm. The best result of each Re-ID algorithm is marked in bold numbers. The best result of the algorithms overall is marked with an asterisk(∗).

algorithms are better than the RRSS and TED method by around 5% on rank one matching accuracy. RRSS and TED have a similar optimization target to our algorithm but without pairwise constraint. This implies that our method is much suitable for Re-ID applications as a result of introducing the pairwise constraint. 3) The performance of the kernelized methods is better than the performance of the linear methods with our algorithm. This is consistent with the mathematical analysis in [85] that kernelization produces more discriminative representation by mapping data into high-dimensional feature space. 4) The active learning algorithms with XQDA method for report better rank one matching accuracy than those with LOMO features.



(a) CAVIAR

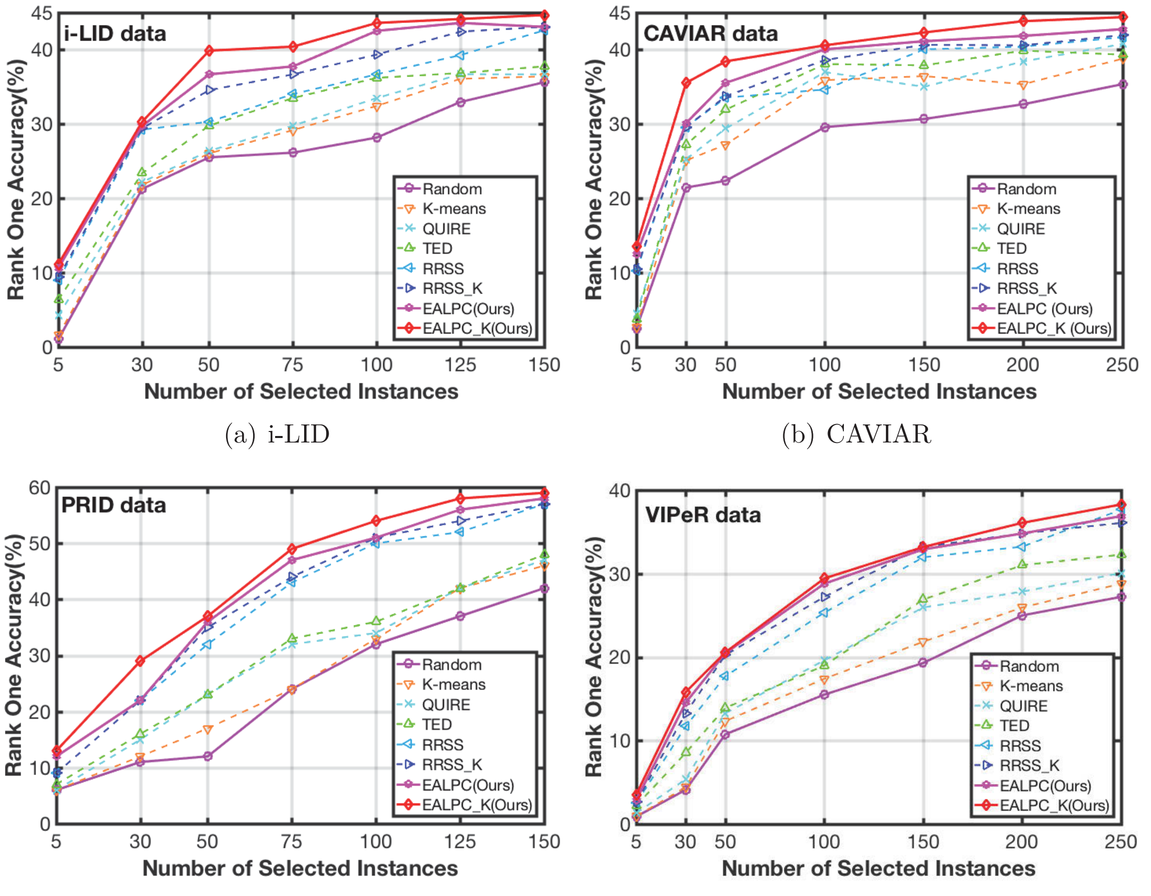
(b) VIPeR

(c) PRID

(d) i-LID

Figure 2.2 CMC Performance Comparison of Active Learning algorithms. XQDA is chosen as the Re-ID algorithm. The percentage of selected samples is set to 10% of all samples.

In Figure 2.2, we illustrate the performance via CMC curves of active learning methods with XQDA as the Re-ID algorithm. The percentage of the labeled training sample is set to only 10% to present a more challenging task. We choose XQDA as it returned the best Re-ID results in the previous experiments. As shown in Fig. 2.2, we can observe that: 1) Our algorithms outperforms other algorithms consistently on all four benchmark datasets. 2) Compared to the results in Table 2.1, all algorithms suffer a decrease in the rank one matching accuracy when the percentage of labeled samples is halved from 20% to 10%. However, our algorithm only decreases around by 5% on rank one matching accuracy whereas the accuracy of others, e.g. Random and K-means, reduces approximately 10%. This indicates that our algorithm is more robust. 3) The matching accuracy of our algorithm is the only one to reach 90% with rank 15 on CAVIAR and VIPeR, and the only one to reach 90% on rank 20 on PRID and i-LID. This implies that our algorithm is more effective on Re-ID .



(a) i-LID

(b) CAVIAR

(c) PRID

(d) VIPeR

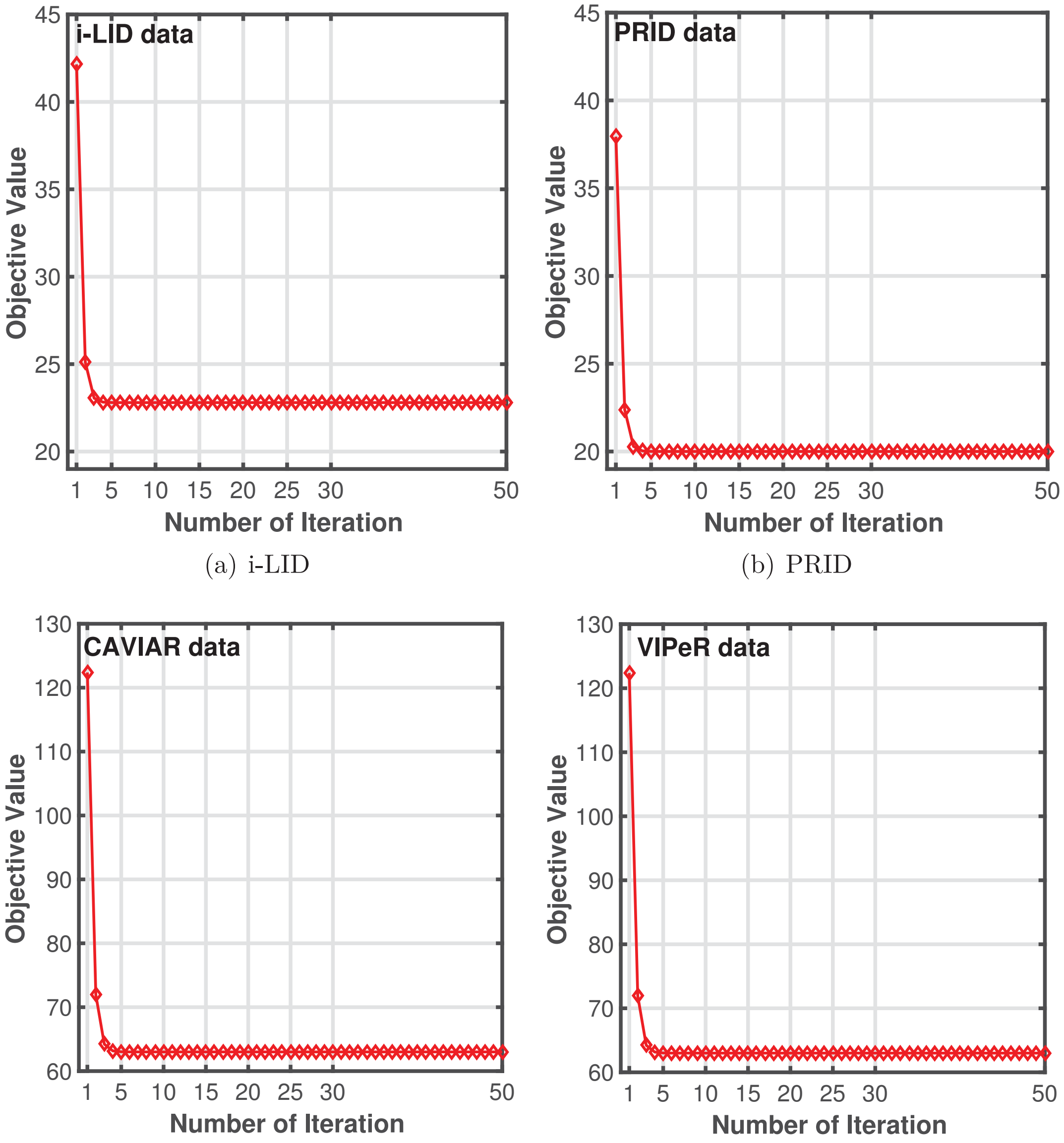
Figure 2.3 Rank One Matching Accuracy(%) w.r.t. Number of Selected Instances. We use XQDA as the Re-ID algorithm and train it with varying numbers of samples selected by the active learning methods.

2). Effects on the Number of Selected Instances. Figure 2.3 illustrates the performance of Re-ID when the number of instances that selected by active learning methods varies. As displayed in Fig. 2.3, we observe that: 1) Generally, rank one matching accuracy of all Re-ID algorithms increases gradually when the number of selected instances increases. 2) All active learning methods report better performances than Random selection. This indicates that active learning algorithms can improve the performance of Re-ID applications. 3) Our algorithm consistently performs better than the other active learning algorithms when the number of selected instance increases. More specifically, for our algorithm, kernelized methods is better than the linear methods.

3). Convergence. In Figure 2.4, we draw the objective value of the first 50 iterations of our algorithm on benchmark datasets. In the experiments, we fix the leverage parameters as $\alpha = 0.1$ and $\beta = 1$ and set the percentage of selected samples to 20%. As shown in Fig. 2.4, the object values of our algorithm decrease dramatically and barely change after the first five iterations on all the benchmark datasets. This indicates that our algorithm converges very rapidly on all the datasets, which is consistent with our theoretical analysis of convergence.

2.5 Summary

In this chapter, we have proposed a novel early active learning algorithm with a pairwise constraint for person re-identification. The proposed method is designed for the early stage of supervised Re-ID experiments when there are limited labor resources for labeling data. Our algorithm introduces a pairwise constant for analyzing graph structures specifically for re-identification. A closed form solution is provided to efficiently weight and select the candidate samples. Extensive experimental studies on four benchmark datasets validate the effectiveness of the proposed algorithm. The experimental results demonstrate that our methods achieve encouraging performance against the state-of-the art algorithms in the filed of early active learning for person re-identification. In future work, our algorithm can be applied to other applications that consider the pairwise relatedness, such as in social network analysis, etc.



(a) i-LID

(b) PRID

(c) CAVIAR

(d) VIPeR

Figure 2.4 Convergence Analysis of EALPC on Benchmark Datasets. The parameters are set as $\alpha = 0.1$ and $\beta = 1$. The percentage of selected samples is 20%.

Chapter 3

Pair-based Early Active Learning

3.1 Background

In last chapter,we have proposed an early active learning algorithm for Re-ID and have achieved promising results in other fields. However,the previous works considers only instance-based active learning schema, but not considers to select the pairs of the samples for labeling, which is essential and important in the person re-identification problem.

In the previous works [72, 85, 130], all of the EAL algorithms focus on finding out the most representative samples on behalf of the whole dataset. However, these algorithms have two drawbacks. First, they are instance-based active learning algorithms that only consider to select instances for annotation. They fail to consider to directly select pairs of samples for annotation, which is essential for Re-ID task. Second, the previous EAL works only consider to select the most representative instances from the whole set of data. They fails to analyze the uncertainty of the samples, which is essential in traditional active learning schema [100]. For Re-ID task, select pairs with higher uncertainty, such as person images with occlusion, will bring more information and enhance the robustness of the Re-ID models.

To overcome the limitations described above, we propose a novel algorithm for person re-identification, Early Active Learning with Pairwise Diversity Maximization, abbreviated as EAL-PDM. The main contributions of our work are as follows:

- To the best of our knowledge we make the first attempt to consider pair-based active learning and simultaneously optimize pairwise uncertainty and diversity of samples for person Re-ID in an early active learning task.

- We are the first to define the person re-identification probability especially for the pairwise uncertainty estimation in EAL for Re-ID task.
- We introduce a pairwise diversity maximization criterion to enhance the pairwise diversity of selected pairs of samples. It is the first attempt that considers pairwise diversity in EAL.
- We propose an efficient algorithm to optimize the proposed object function. Our optimization algorithm also provides guarantees to reach the global optimum in the convergence.

3.2 The Proposed Framework

In this section, we first revisit the early active learning. Then, we propose our early active learning with pairwise diversity maximization for Re-ID .

3.2.1 Pairwise Uncertainty and Diversity

In this section, we focus on pair-based early active learning for the person Re-ID problem. Person Re-ID can be formed as an image *retrieval* task which aims to re-identify the same person across non-overlapping camera views. Therefore, the analysis of pairwise relativeness of images across different camera views is essential and important. In active learning stage, it is desired to select the most informative and diverse pairs of samples to generate training data to enhance the performance of Re-ID methods. We evaluate the uncertainty and diversity of pairs of samples as follows:

Pairwise Uncertainty Estimation. To select the most informative subset of pairs of samples, a common criterion is via uncertainty estimation. In this chapter, we prefer to evaluate the pairwise entropy of samples to benefit the Re-ID methods rather than the instance entropy. In order to achieve this, we first define the *person re-identification probability* as

$$p(l_{ij}|\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1 + \exp\{\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathcal{M}}^2 - \eta\}}, \tag{3.1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ is a pair of samples and l_{ij} is the pairwise label. The person re-identification probability estimates how likely $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same person (i.e., $l_{ij} = 1$) or not (i.e., $l_{ij} = 0$). The positive semi-definite matrix $\mathcal{M}$ is a pre-defined metric and $\eta > 0$ is a threshold. Accordingly, $\mathbf{x}_i$ and $\mathbf{x}_j$ are more likely to belong to the same person when $\|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathcal{M}} \le \eta$. In this chapter we simply let $\mathcal{M} = \mathbf{I}$. The discussion of learning $\mathcal{M}$ is left for specific research topics such as metric learning [121]. Overall, the uncertainty of samples is estimated by the pairwise entropy defined as:

$$H(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{2}\sum_{i,j} p(l_{ij}|\mathbf{x}_i, \mathbf{x}_j) \log p(l_{ij}|\mathbf{x}_i, \mathbf{x}_j). \tag{3.2}$$

Pairwise Diversity Maximization. In order to maximize the diversity of the selected samples, we aim to reduce the number of pairs where samples are very similar to each other. To evaluate the similarity, a similarity matrix $\mathbf{K}$ is introduced. Each element of $\mathbf{K}$ can be defined as $k_{i,j} = -\|\mathbf{x}_i - \mathbf{x}_j\|/\sigma^2$ with σ the parameter. We define $\Omega(\mathbf{u}_i) = \frac{1}{2}\sum_{i,j} u_i u_j k_{ij} = \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{K}\mathbf{u}$. Given two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, $k_{i,j}$ will be larger if they are more similar to each other. Therefore, minimize $\Omega(\mathbf{u})$ will enforce the two samples not to be selected simultaneously (i.e, either u_i or u_j or both of them are forced to zero).

3.2.2 Pair-based Early Active Learning

Finally, we propose the pair-based early active learning schema. Unlike the former works [72] which provide a relaxation of the problem, we reformulate the EAL problem to an equivalent problem that can be solved efficiently. As mentioned before, learning the optimal subset in Eq. (3.3) is NP-hard. We introduce a indicator vector $\mathbf{u} \in \mathbb{R}^{n\times 1}$ with $u_i \in \{0,1\}$. The sample selected is assigned by $u_i = 1$ otherwise $u_i = 0$. Therefore, we can represent $\mathbf{V} = \mathbf{X}\text{diag}(\mathbf{u})$. Then, taken uncertainty estimation in consideration, the optimization problem of early active learning is reformulated as:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{u}} \mathbf{R}(\mathbf{X}, \mathbf{A}, \mathbf{u}) + \alpha\mathbf{I}(\mathbf{X}, \mathbf{u}) + \Omega(\mathbf{u}), \\ &\quad s.t.\ \mathbf{u}^\mathsf{T}\mathbf{1} = m,\ u_i \in \{0,1\}. \end{aligned} \tag{3.3}$$

where the first term $\mathbf{R}(\mathbf{X}, \mathbf{A}, \mathbf{u})$ is the representative learning function and $\mathbf{I}(\mathbf{X}, \mathbf{u})$ is the informative learning function with a leverage parameter $\alpha > 0$. We further introduce a diversity maximization term $\Omega(\mathbf{u})$ which will be discussed later. After introducing $\mathbf{u}$, we reformulate Eq. (2.2) to

$$\mathbf{R}(\mathbf{X}, \mathbf{A}, \mathbf{u}) = \|\mathbf{X} - \mathbf{X}\text{diag}(\mathbf{u})\mathbf{A}\|^2_F. \tag{3.4}$$

Meanwhile, we can formulate the uncertainty objective function as

$$\mathbf{I}(\mathbf{X},\mathbf{u}) = \sum_{i}^{n} u_i(\sum_{j\neq i}^{n} -H(\mathbf{x}_i,\mathbf{x}_j)) = \mathbf{u}^\mathsf{T}\xi, \tag{3.5}$$

where for each element $\xi_i \in \xi$, we calculate $\xi_i = \sum_{j\neq i}^{n} -H(\mathbf{x}_i,\mathbf{x}_j)$ which estimates the sum of pairwise uncertainty between $\mathbf{x}_i$ and the rest of samples $\mathbf{x}_j, j \neq i$. Notice we want to choose the pairs with high uncertainty, such that we minimize the negative entropy in the objective function.

Finally, the overall optimization problem of the proposed algorithm becomes to:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{u}} \|\mathbf{X} - \mathbf{X}\mathrm{diag}(\mathbf{u})\mathbf{A}\|_F^2 + \alpha\mathbf{u}^\mathsf{T}\xi + \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{K}\mathbf{u}, \\ &s.t.\ \mathbf{u}^\mathsf{T}\mathbf{1} = m,\ u_i \in \{0,1\}. \end{aligned} \tag{3.6}$$

3.2.3 Optimization

In this section we propose a effective algorithm optimize the active learning problem formulated in Eq. (3.6). Note solve Eq. (3.6) with u_i are integer is hard, we relax the problem to

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{u}} \|\mathbf{X} - \mathbf{X}\mathrm{diag}(\mathbf{u})\mathbf{A}\|_F^2 + \alpha\mathbf{u}^\mathsf{T}\xi + \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{K}\mathbf{u}, \\ &s.t.\ \mathbf{u}^\mathsf{T}\mathbf{1} = 1, u_i \geq 0. \end{aligned} \tag{3.7}$$

The objective function in Eq. (3.7) can be solved by alternately optimize $\mathbf{A}$ and $\mathbf{u}$. After getting the optimized continues variable of $\mathbf{u}^*$, the optimal indicator vector can be get by a truncate function $\mathbf{u} = truncate(\mathbf{u}^*, m)$ which let the top m larger u_i in $\mathbf{u}$ be ones and the rest be zeros. Next we discuss the optimization of Eq. (3.7).

Optimize u with fixed A. With $\mathbf{A}$ fixed, solve the objective function in Eq. (3.7) w.r.t. $\mathbf{u}$ is a standard quadratic programming (QP) problem. We propose to solve it by an algorithm based on the augmented Lagrange multiplier (ALM) framework [17]. First we rewrite Eq. (3.7) as follows:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{u}} \|\hat{\mathbf{x}} - \mathbf{Q}\mathbf{u}\|_2^2 + \alpha\mathbf{u}^\mathsf{T}\xi + \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{K}\mathbf{u}, \\ &s.t.\ \mathbf{u}^\mathsf{T}\mathbf{1} = 1, \mathbf{u} = \mathbf{p}, \mathbf{p} \geq 0, \end{aligned} \tag{3.8}$$

where $\widehat{\mathbf{x}} = \mathrm{vec}(\mathbf{X})$ and $\mathbf{Q} = [\mathrm{vec}(\mathbf{q}_1), \cdots, \mathrm{vec}(\mathbf{q}_n)]$ where $\mathbf{q}_i = \mathbf{x}_i\mathbf{a}^i$. The augmented Lagrangian function of Eq. (3.8) is formulated as

$$\begin{aligned} L(\mathbf{u},\mathbf{p},v,\lambda_1,\lambda_2) &= \frac{v}{2}(\mathbf{u}^\mathsf{T}\mathbf{1}-1+\frac{1}{v}\lambda_1)^2 \\ +\frac{v}{2}\|\mathbf{u}-\mathbf{p}+\frac{1}{v}\lambda_2\mathbf{1}\|_F^2 &+ \|\widehat{\mathbf{x}}-\mathbf{Q}\mathbf{u}\|_2^2+\alpha\mathbf{u}^\mathsf{T}\xi+\frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{K}\mathbf{u}, \\ s.t.\ \mathbf{p} &\geq 0. \end{aligned} \tag{3.9}$$

Then we can optimize $\mathbf{u}$ with fixed $\mathbf{p} = \mathbf{p}^*$ from

$$\min_{\mathbf{u}} L(\mathbf{u},\mathbf{p}^*,v,\lambda_1,\lambda_2) \Leftrightarrow \min_{\mathbf{u}} \mathbf{u}^\mathsf{T}\mathbf{b}+\frac{1}{2}\mathbf{u}^\mathsf{T}\Sigma\mathbf{u}, \tag{3.10}$$

where $\Sigma = \mathbf{K}+v\mathbf{I}+v\mathbf{1}\mathbf{1}^\mathsf{T}+2\mathbf{Q}^\mathsf{T}\mathbf{Q}$ and $\mathbf{b} = (\lambda_1\mathbf{1}-v\mathbf{1})+(\lambda_2\mathbf{1}-v\mathbf{p}^*)+\alpha\xi-2\mathbf{Q}^\mathsf{T}\widehat{\mathbf{x}}$. Thus the optimal solution of $\mathbf{u}$ is

$$\mathbf{u}^* = \Sigma^{-1}\mathbf{b}. \tag{3.11}$$

After determined $\mathbf{u} = \mathbf{u}^*$, we can optimize $\mathbf{p}$ from

$$\min_{\mathbf{p}\geq 0} L(\mathbf{u}^*,\mathbf{p},v,\lambda_1,\lambda_2) \Leftrightarrow \min_{\mathbf{p}\geq 0} \|\mathbf{p}-(\mathbf{u}+\frac{1}{v}\lambda_2)\|_F^2. \tag{3.12}$$

By solving the above optimization problem, the solution of $\mathbf{p}$ is

$$\mathbf{p}^* = \mathrm{pos}(\mathbf{u}+\frac{1}{v}\lambda_2\mathbf{1},0). \tag{3.13}$$

where $\mathrm{pos}(\mathbf{z},\epsilon)$ is a truncate function which assigns zero to each element of $\mathbf{z}$ less than ϵ, i.e., $\forall z_i \in \mathbf{z}, \mathrm{pos}(z_i) = \max(z_i,\epsilon)$.

Optimize A with fixed u. Optimize $\mathbf{A}$ in Eq. (3.6) with $\mathbf{u}$ fixed can be calculated by a closed form solution.

$$\mathbf{A}^* = (\mathbf{V}^\mathsf{T}\mathbf{V})^{-1}\mathbf{V}\mathbf{X}, \tag{3.14}$$

where $\mathbf{V} = \mathbf{X}\mathrm{diag}(\mathbf{u}^*)$.The overall optimization algorithm for solving $\mathbf{u}$ is described in Algorithm 1. It can be verified that Algorithm 1 converges to the *global optimum*.

```
Algorithm 2: Algorithm for solving problem in Eq. (3.6)
Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, parameters $\alpha$, $v$
Initialize $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\forall u_i \in \mathbf{u}, u_i = \frac{1}{n}$. Set $\mathbf{p} = \mathbf{u}$, $\lambda_1 = 0$ and $\lambda_2 = 0$.
while not converge do
    Update $\Sigma$ by $\Sigma = \mathbf{K} + v\mathbf{I} + v\mathbf{1}\mathbf{1}^\mathsf{T} + 2\mathbf{Q}^\mathsf{T}\mathbf{Q}$.
    Update $\mathbf{b}$ by $\mathbf{b} = (\lambda_1\mathbf{1} - v\mathbf{1}) + (\lambda_2\mathbf{1} - v\mathbf{p}^*) + \alpha\xi - 2\mathbf{Q}^\mathsf{T}\widehat{\mathbf{x}}$.
    Compute $\mathbf{u}^*$ by solving the linear system $\Sigma\mathbf{u} = \mathbf{b}$.
    Compute $\mathbf{p}^*$ by $\mathbf{p}^* = \mathrm{pos}(\mathbf{u}^* + \frac{1}{v}\lambda_2)$.
    Compute $\mathbf{A}^* = (\mathbf{V}^\mathsf{T}\mathbf{V})^{-1}\mathbf{V}\mathbf{X}$.
    Update $\lambda_1$ by $\lambda_1 = \lambda_1 + v \times (\sum_{i=1}^{n} u_i - 1$.
    Update $\lambda_2$ by $\lambda_2 = \lambda_2 + v \times (\mathbf{u} - \mathbf{p})$.
    Update $v = \rho v$.
Output: $\mathbf{u}^*$.
```

3.3 Experimental Study

3.3.1 Experimental Settings

Datasets

Four widely referred Re-ID benchmark datasets are utilized in the experiments.

1. The **VIPeR** [30] dataset collects camera shots of 632 persons from two non-overlapping camera views. The total number of images are 1,264 and for each person there are one image captured from each different camera. variations of viewpoint and illumination condition appears frequently in the dataset.

2. The **PRID** [35] dataset captures camera shots of 385 individuals from two distinct cameras. One of the cameras collects images of 749 persons and the other one collects images of 385 person. There are 200 person being captured in both of the cameras.

3. The **i-LID** [139] dataset contains camera shots of 119 individuals in an airport terminal. There are 476 images images captured by three non-overlapping cameras. Large occlusions are observed because of the carry-on luggages and viewpoint variations.

4. The **CAVIAR** [14] dataset records images of 72 individuals in a shopping mall. A total number of 1,220 images are captured by two different cameras. There are 10 to 20 images for each individual. In the dataset, size of the camera shots varies significantly from 39×17 to 141×72.

Dataset	CAVIAR					VIPeR				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	23.65	23.47	21.38	27.55	25.87	26.65	23.01	27.23	22.78	23.64
K-means	26.90	25.99	22.05	27.74	27.40	27.59	26.16	27.59	23.15	24.39
TED	29.78	28.70	29,42	27.94	28.08	27.45	28.53	28.43	25.75	26.09
QUIRE	30.66	30.87	31.56	28.18	26.16	28.39	27.43	28.54	26.25	25.13
RRSS	31.87	30.69	33.57	30.95	29.01	31.56	28.54	30.71	27.34	28.04
RRSS_K	31.69	33.03	35.56	31.41	31.13	31.61	28.73	31.46	28.51	29.40
EALPC	34.12	33.57	37.45	33.09	31.16	32.61	29.45	31.82	28.54	29.56
EALPC_K	35.00	35.20	38.75	33.29	31.91	33.66	30.44	34.29	29.18	30.03
EAL-PDM	**36.29**	**37.13**	**39.55***	**34.22**	**32.25**	**34.00**	**31.24**	**36.53***	**30.02**	**30.53**
Dataset	PRID					iLIDS				
Algorithm	NFST	KCCA	XQDA	kLFDA	MFA	NFST	KCCA	XQDA	kLFDA	MFA
Random	24.49	25.47	24.00	23.50	20.00	25.96	23.40	25.00	23.35	25.00
K-means	26.16	27.54	27.01	24.70	21.20	27.02	23.94	27.00	25.57	25.20
TED	27.72	27.71	29.32	24.33	22.11	29.15	25.33	28.13	27.33	29.20
QUIRE	27.24	26.90	29.33	24.40	22.50	28.72	25.74	28.03	29.48	30.20
RRSS	29.21	28.44	30.00	25.09	23.97	28.11	27.66	30.82	30.08	30.55
RRSS_K	30.33	29.03	31.05	25.30	24.10	29.17	27.37	32.00	30.30	31.10
EALPC	32.22	30.63	31.03	25.90	25.60	29.26	27.66	32.34	30.43	31.60
EALPC_K	32.70	31.50	33.40	26.06	25.70	31.19	28.72	34.00	31.60	32.47
EAL-PDM	**33.08**	**33.95**	**35.27***	**27.10**	**25.91**	**35.39**	**29.25**	**36.22***	**33.42**	**33.14**

Table 3.1 Rank One Matching Accuracy(%) on the benchmark datasets. 20% of the samples are selected for labeling. The rows are Re-ID algorithms. The columns are active learning algorithms. For each Re-ID methods, the best result w.r.t active learning algorithms is marked in bold. The best result in all the Re-ID methods is marked with an asterisk($*$).

In the pre-processing stage, Local Maximal Occurrence (LOMO) features for person image representation [62] are implemented for person images as in previous works [69, 89]. After rescaling all images to 128 × 48 pixels, we follow the default setting in [62] to produce LOMO features.

Active Learning Algorithms

Six active learning algorithms are compared with the proposed algorithm (i.e., **EAL-PDM**) in the experiments.

1. **Random:** We randomly select samples for labeling as a baseline algorithm.

2. **K-means:** As suggested in [85], we apply the K-means algorithm as another baseline. Rank of the samples are sorted by their distances to the nearest cluster centers.

3. **QUIRE:** [40] Active learning by Querying Informative and Representative Examples (QUIRE) is an algorithm which use a min-max margin-based approach to query the most informative and representative samples. However, it require a set of pre-labeled samples to verify the importance of the unlabeled samples.

4. **TED:** [130] Active Learning via Transduction Experimental Design (TED) is an algorithm is an early active learning method which select the most representative subset of samples from the unlabeled dataset. It formulates the subset selection task as a regularized linear regression problem.

5. **RRSS:** [85] Early active learning via Robust Representation and Structured Sparsity (RRSS) is another early active learning algorithm. It introduces a $\ell_{2,1}$-norm regularization term to TED for robustness and structure sparsity. In their work, authors also propose a kernelized algorithm **RRSS_K** with a Gaussian kernel.

6. **EALPC:** [72] Early Active Learning with Pairwise Constraint (EALPC) is a recently proposed algorithm. It introduces a pairwise constraint to EAL in order to force the similar representations to be close to each other. It also provides a kernelized version **EALPC_K** with a Gaussian kernel.

Person Re-identification Algorithms

Five supervised Re-ID methods are applied to evaluate the performance of the compared active learning algorithms on the Re-ID task.

1. **NFST:** [132] Null Foley-Sammon Transform space learning (NFST) is a Re-ID method which learns a discriminative subspace, where samples belonging to the same person are collapsed to a single point.
2. **KCCA:** [69] Kernel Canonical Correlation Analysis (KCCA)method aims to learn a common subspace of the images of person from the non-overlapping cameras.
3. **XQDA:** [62] Cross-view Quadratic Discriminant Analysis (XQDA)seeks a discriminant low dimensional subspace by cross-view quadratic discriminant analysis and applies metric learning in the space.
4. **kLFDA:** [122] Kernelized Local Fisher Discriminant Classifier (kLFDA) applies a kernelized method with closed form solution to handle large dimensional feature vectors by maximizing a Fischer optimization criterion.
5. **MFA:** [126] Marginal Fisher Analysis (MFA) learns graphs that characterize the intra-class compactness and interclass separability and applies them for dimensionality reduction.

Settings

In the experiments, as advised in previous works [72, 89], we randomly splits each dataset into two subsets with equal number of samples for training and testing. There are no overlapping of persons in the training and the testing sets. To generate the probe and gallery sets for Re-ID task, images captured from one of the camera views are chosen as the probe sets and the remaining images from the rest camera views are assigned as gallery sets. Overall, we randomly and independently split each of the datasets 10 times to generate 10 trails. The average performance of all the trails are report as the final result.

The procedure of the experiments is as follows: first, we apply active learning methods to select a certain number of samples and query human annotator to label them. Then, the supervised Re-ID methods are applied to train the Re-ID models. Finally, we evaluate the Re-ID performance on the testing sets. We measure the performance of Re-ID models by Cumulative Matching Characteristic (CMC) curve, which is commonly introduced in Re-ID works [47, 62, 68, 72]. Specifically, CMC computes the cumulative probability that a image in the top k gallery set happens to match the probe image. In the experimental study, we report the Rank One (CMC R1) matching accuracy from CMC for simplicity. For each algorithms, we run a ten-fold cross validation strategy to grid search the optimal parameters (if any) in a region of $\{10^{-4}, 10^{-3}, \cdots, 1, \cdots, 10^{3}, 10^{4}\}$.

3.3.2 Results Analysis

Performance of Active Learning for Re-id

To investigate the performance of active learning for Re-ID , we employ the active learning methods to select 20% of total data for labeling on four benchmark datasets. Then, the Re-ID methods are trained with the selected and labeled data. We illustrate the testing result in Table 3.1 in rank one matching accuracy. In the table, active learning algorithms are displayed by rows and supervised Re-ID methods are displayed by columns. In Table 3.1, we observe that:

1) All of the active learning algorithms achieve higher performance compared to the Random selection algorithm consistently on the four datasets. It verifies that the active learning algorithms can select contributive samples to improve the performance of the Re-ID methods.

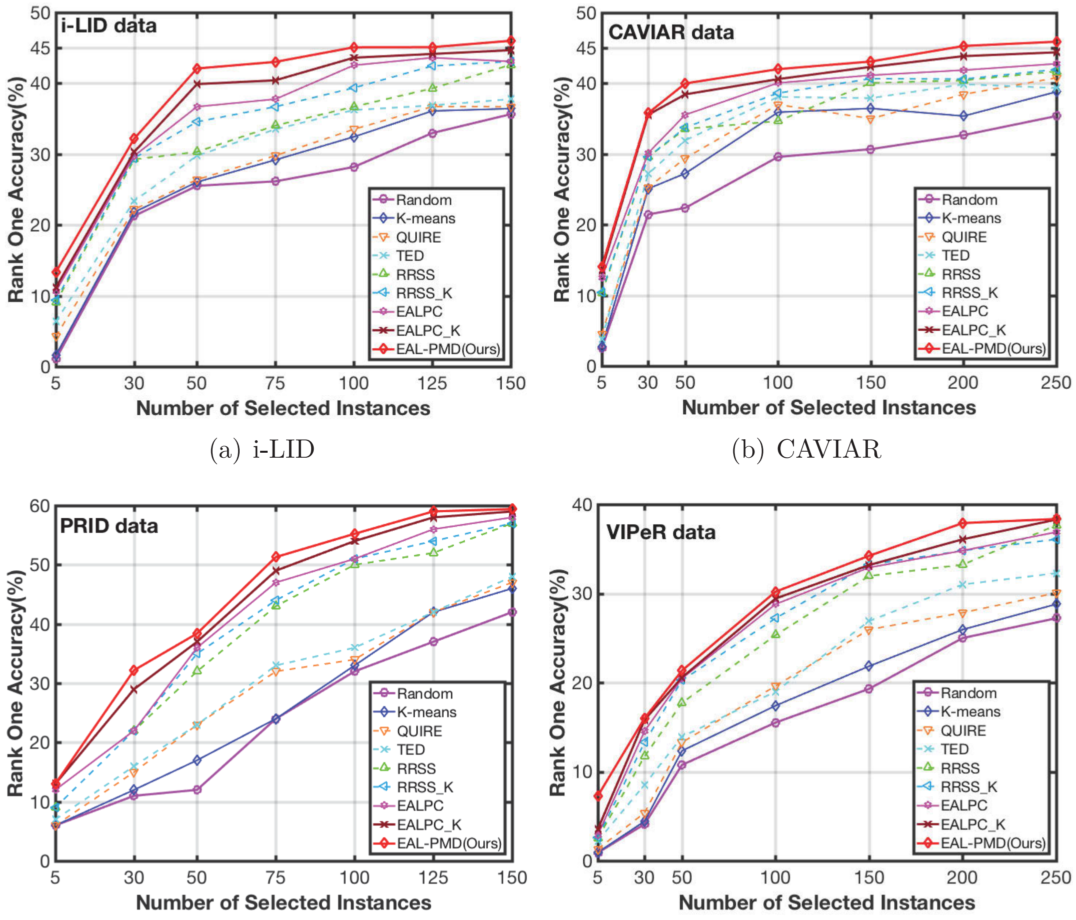
2) For EAL algorithms, our algorithm EAL-PDM and EALPC, which is specified for Re-ID , outperforms the former EAL algorithms RRSS and TED at around 5% to 7%. It implies that the EAL algorithm that consider the pairwise relationships will enhance the active learning performance for the Re-ID task.

3) More over, our algorithm EAL-PDM consistently reports higher rank one matching accuracy than EALPC at around 2% on all of the four datasets. It further indicates that the pair-based early active learning can enhance the Re-ID task and get better performance than the instance-based algorithms EALPC.

Influence of the Number of Selected Instances

In order to investigate the influence of the number of selected instances, we record the rank one matching accuracy while increasing the number of selected instance. We employ XQDA on behalf of the Re-ID methods as it reports the best performance in Table 3.1. In Figure 3.1. Specifically, for the convenience of comparisons, for our algorithm, we also record the number of instances rather than the number of pairs. In the experiments, we observe that:

1) All of the active learning algorithms are outperform the baseline algorithm, i.e., Random selection. It confirms that even the number of selected samples are very small (e.g.,5), the active learning algorithms can select contributive samples to enhance the Re-ID methods.



(a) i-LID

(b) CAVIAR

(c) PRID

(d) VIPeR

Figure 3.1 Rank One Matching Accuracy(%) w.r.t. Number of Selected Samples. XQDA is chosen as the Re-ID algorithm and trained with samples determined by the active learning methods with different amounts.

2) All of the Re-ID algorithms are gradually improved when the number of selected instance increases. It indicates that increasing the amount of useful labeled data will improve the performance of the Re-ID task.

3) Our algorithm EAL-PDM consistently outperforms the rest of active learning algorithms when the number of selected instance varies. More specifically, our algorithm distinctly outperforms the rest algorithm since more than 50 instances are selected. It implies our algorithm can efficiently select useful pairs of samples to enhance the Re-ID tasks.

Convergence

In order to verify the convergence, we force our algorithm to run 50 iterations and record the object values on the benchmark datasets. We illustrate the performance of CAVIAR dataset in Figure 3.2 and the performance on the other three datasets is very similar. We fix $\alpha = 0.1$ and set the percentage of selected samples to 20% in the experiment. It can be observed that the object values decrease very quickly in around first 10 iterations and barely changes after that. It implies that our algorithm converges very fast.

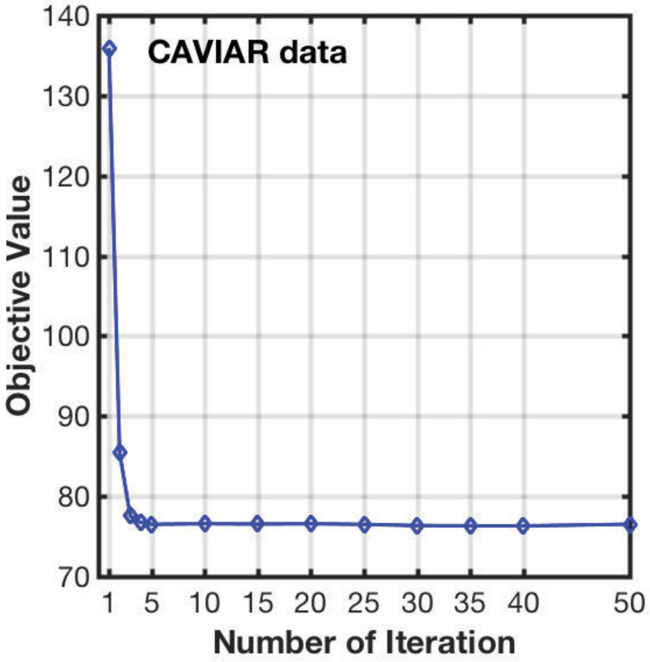


Figure 3.2 Convergence of the Proposed Algorithm on CAVIAR Dataset

Diversity Analysis

We investigate the diversity of our algorithms compared to EALPC on dataset i-LID, both of which aims to analyze pairwise relativeness in Re-ID active learning. We evaluate the diversity via calculating the average pairwise similarity (APS) defined as $\frac{1}{2S}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{V},i\neq j}\exp\{-\|\mathbf{x}_i-\mathbf{x}_j\|/\sigma^2\}$ where $\sigma=2$ is a scale parameter. S is the number of pairs constructed by the samples $\mathbf{V}$ determined by the active learning algorithms. Ideally, higher APS indicates the similarity of selected pairs are averagely large. As shown in Figure 3.3, we can observe that the average pairwise similarity of our algorithm is consistently lower than EALPC. We number of selected instances are lower than 30, the APS of EALPC is very high. It indicates that the instance-based algorithm cannot select diverse samples. On the contrary, our algorithm consistently report lower APS (around 0.8) in the experiments even when the number of selected instances are very few. It indicates that our algorithms could select less similar pairs which improves the diversity of the selected data.

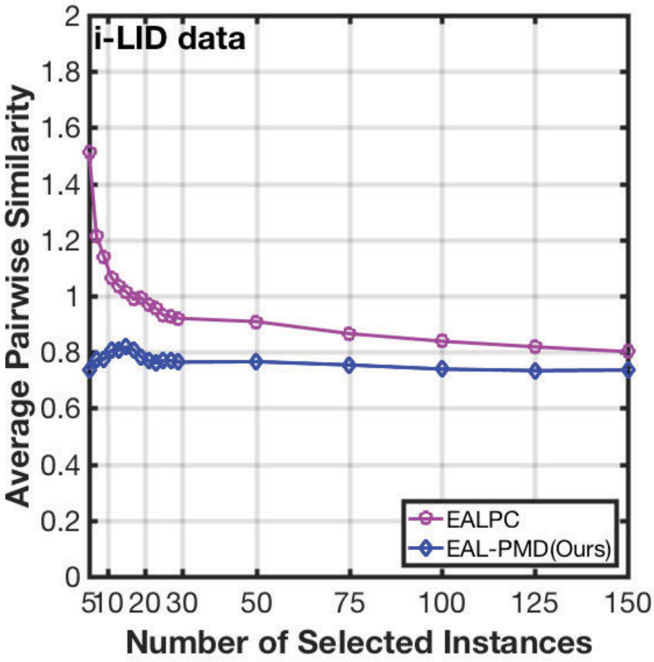


Figure 3.3 Diversity Analysis of the Proposed Algorithm.The leverage parameter is $\alpha=0.1$ and 20% of the samples are selected for annotation.

3.4 Summary

In this chapter, we have proposed a new early active learning algorithm with pairwise diversity maximization for person re-identification, i.e., EAL-PDM. The proposed algorithm considers both uncertainty and representativeness in EAL. Moreover, with pairwise diversity maximization, it improve the diversity of the selected data by reducing the chance of select similar instance pairs. Experimental study on four benchmark datasets demonstrates the superior performance of our algorithm over several state-of-the-art active learning algorithms in Re-ID tasks. For future works, we suggest applying our algorithm to our works depended on the pairwise relatedness, such as social network analysis, etc.

Chapter 4

Semi-supervised Bayesian Attribute Learning

4.1 Background

Existing approaches to Re-ID have mainly focused on representation learning and/or metric learning to overcome these challenges. In representation learning, many frameworks learn a factor-based representation to enhance Re-ID task performance [50, 73]. Several recent works have turned to attribute learning methods for further improvement [66]. In Re-ID, attributes are mid-level features shared by multiple instances, such as hair color or wearing/not wearing a dress. Overall, the general idea behind these methods is that there are only a certain number of feature subsets that contribute to image matching performance. In metric learning, algorithms learn a suitable metric in the given set of data, which is then used to measure similarity [36, 122].

Previous studies have demonstrated some exciting results, but there are still challenges associated with each approach. Determining the number of latent factors in factor-based representation models is a common problem. Typically, cross-validation forms the solution, where the model evaluates various numbers of latent factors that are manually pre-defined. However, this is a time-consuming task for large Re-ID datasets, limiting the scalability of these methods. Another solution is to manually annotate the attributes to enhance learning performance [66, 104]. However, on large-scale datasets, this method has high human labor costs.

Metric learning Re-ID methods also have drawbacks. Because they rely on learning a metric suitable to the Re-ID targets, the performance is sensitive to the given dataset. Additionally, choosing the optimal method for calculating similarity distances, e.g., using ℓ_1 norm or ℓ_2 norm, can be problematic [114]. Moreover, metric learning methods rely on pair-wise label information, such that performance suffers when there are only a few labels [127].

A few previous works have attempted to jointly apply both representation and metric learning to Re-ID problems. Some use each technique independently to accomplish a specific goal. For instance, in [62, 132], researchers use representation learning methods in pre-processing stage to generate useful features that can then be used in metric learning. Others combine both methods into a deep learning architecture. However, these methods still rely on pre-annotated attributes [66, 104] or labeled data [1].

To overcome these limitations with Re-ID tasks, we propose a semi-supervised Bayesian attribute learning algorithm (SBAL). SBAL combines an Indian buffet process (IBP) [27] prior in an infinite latent factor model that enables adaptively learning attributes for Re-ID [6]. Additionally, inspired by statistical relation learning, we also propose Re-ID probability, which has been successfully used in knowledge graph learning on large-scale datasets, such as social networks [83]. Wrapped within a Bayesian framework, SBAL automatically determines the latent factors and simultaneously estimates a *re-identification probability*. The contributions of our work are as follows:

- We introduce IBP as the prior of latent factors for learning binary representations. A dictionary of attributes is adaptively determined using an efficient estimation method. Thus, our algorithm does not require the dimensionality of latent factors to be pre-defined, nor the attribute information to be pre-annotated for training, which are two major limitations of the existing frameworks.

- We propose a re-identification probability for predicting pair-wise relations in Re-ID . The Re-ID probability does not rely on distance computation and avoids the problem of determining the optimal method for computing distances inherent in traditional metric learning.

- We propose a Bayesian framework unifies representation learning and Re-ID probability estimation and can simultaneously optimize both learning tasks.

- Our algorithm is also able to estimate unknown pair-wise labels using the probability distributions learned from known pair-wise labels, making our algorithm robust in semi-supervised learning scenarios.

4.2 The Proposed Method

4.2.1 Formulation

The following section formally describes each component of the proposed framework. This section focuses on two-view Re-ID problems, where data is recorded from two non-overlapping camera views. However, the schema is easy to extend to multi-view Re-ID scenarios.

Let $\mathbf{X} = \{\mathbf{X}_1; \mathbf{X}_2\}$ be the training data, in which $\mathbf{X}_1 \in \mathbb{R}^{d\times n_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{d\times n_2}$ are image sets of people from two non-overlapping cameras, camera 1 and 2, that containing n_1 and n_2 images respectively. We also have a matrix of pair-wise labels $\mathbf{Y} \in \mathbb{R}^{n_1\times n_2}$, where $\mathbf{y}_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are the same person; otherwise, $\mathbf{y}_{ij} = -1$. However, not all the pairs have labels, i.e., $\mathbf{Y}$ is not fully observed. Let $\mathbf{y}_{ij} = 0$ indicate the unknown pair-wise labels for observations $\mathbf{x}_i$ and $\mathbf{x}_j$. The set of pairs with known labels is denoted as $\mathcal{I} = \{(i,j)|y_{ij} \in \{-1,1\}\}$ and the set of pairs without labels is denoted as $\mathcal{U} = \{(i,j)|y_{ij} = 0\}$.

The first step is to learn representations of the training data with a Bayesian generative model. Let $\mathbf{A} \in \mathbb{R}^{d\times k}$ be a dictionary of basic patterns (attributes) on k basis. Let $\mathbf{Z} \in \mathbb{R}^{k\times n}$ be a binary representation matrix of $\mathbf{X}$ where $z_{ik} \in \{0,1\}$ and $z_{ik} = 1$ indicates the presence of attribute $\mathbf{a}_k$ for the image otherwise $\mathbf{x}_i$ and $z_{ik} = 0$. Given a set of images $\mathbf{X}$ we therefore have $\mathbf{X} \approx \mathbf{AZ}$. After learning a dictionary of attributes $\mathbf{A}$, the binary representation of a new image $\mathbf{x}$ can be obtained by $\mathbf{z} = \arg\min_{\hat{z}\in\{0,1\}} \|\mathbf{x} - \mathbf{A}\hat{z}\|_2^2$. The prior distributions of $\mathbf{A}$ and $\mathbf{X}$ are usually assumed to be Gaussian [6]:

$$\mathrm{P}(\mathbf{A}|0,\sigma_{\mathbf{A}}^2) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{N}(a_{dk};0,\sigma_{\mathbf{A}}^2), \tag{4.1}$$

and

$$\mathrm{P}(\mathbf{X}|\mathbf{Z},\mathbf{A},\sigma_{\mathbf{X}}^2) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{x}_i;\mathbf{A}z_i,\sigma_{\mathbf{X}}^2 I). \tag{4.2}$$

The above formulations assume that the dimensionality K of the latent factor $\mathbf{Z}$ is known as a priori. However, this assumption is often unrealistic in practice, particularly with large-scale datasets, as the possible attributes in image data become more complex when the size of the dataset increases. Conventional methods [50, 60] usually include a model selection stage, such as cross-validation, to select an appropriate value for K by retraining and evaluating the model. This is an expensive process when the training data is large and may even miss the optimal value of K if it is outside the range of the search.

We overcome this problem by introducing Indian Buffet Process (IBP) as the prior of $\mathbf{Z}$. IBP is a nonparametric prior and has been widely used in infinite latent factor models [6, 27]. These models are based on the assumption that an infinite number of latent factors have a distribution using an IBP prior. Considering finite latent factor models first, our model assumes there is a binary feature vector $\mathbf{z}_i$ with K elements for each instance $\mathbf{x}_i$, i.e., $\mathbf{z}_i \in \{0,1\}^K$. Further, we assume $\mathbf{Z}$ has a prior distribution of:

$$\mathrm{P}(\mathbf{Z}|\alpha) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \tag{4.3}$$

The binary latent factor z_{ik} is drawn from a Bernoulli distribution, Bernoulli(π_k), and parameterized by π_k. Furthermore, we assume π_k is sampled from a Beta distribution Beta(α/K, 1) where α is the hyper-parameter and K is the number of basis (i.e. attributes). $m_k = \sum_{i=1}^{N} z_k$ denotes the total number of times the kth attribute in the N samples is found. Then, according to the infinite assumption, i.e. letting $K \to \infty$, we obtain the IBP prior of the binary representations [6]:

$$\lim_{K\to\infty} \mathrm{P}(\mathbf{Z}|\alpha) = \frac{\alpha^{K_+}\exp(-\alpha H_N)}{K_+!} \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}, \tag{4.4}$$

where $H_N = \sum_{i=1}^{N} i^{-1}$ is the Nth harmonic number and K_+ denotes the number of determined attributes corresponding to the dataset $\mathbf{X}$. Several methods for inferring the prior in (4.4) have been proposed in previous works, such as sampling methods and variational methods [27]. However, they can be computationally expensive when the number of instances N becomes large. As a more efficient alternative, we propose learning the joint probability $\mathrm{P}(\mathbf{X},\mathbf{A},\mathbf{Z}) = \mathrm{P}(\mathbf{X})\mathrm{P}(\mathbf{A})\mathrm{P}(\mathbf{Z})\mathrm{P}(\mathbf{X}|\mathbf{Z},\mathbf{A})$ with an asymptotic limitation as in [6]. The details of this approach are provided in the next section.

In the second step, with the binary representations of images, we formalize the Re-ID task as a probabilistic relation learning schema. The *re-identification probability* that the image representations $\mathbf{z}_i$ and $\mathbf{z}_j$ include the same person is calculated by

$$\mathrm{P}(y_{ij} = 1|\mathbf{Z},\mathbf{W}) = \eta(\mathbf{z}_i W \mathbf{z}_j^T), \tag{4.5}$$

where $\mathbf{z}_i \in \mathbf{Z}_1$ and $\mathbf{z}_j \in \mathbf{Z}_2$, $\eta(v) = \frac{1}{1+\exp(-v)}$ is the sigmoid function. We assume the real value matrix $\mathbf{W} \in \mathbb{R}^{K\times K}$ is drawn from a Gaussian prior:

$$\mathrm{P}(\mathbf{W}|\Theta,\sigma_{\mathbf{W}}^2) = \prod_{(k,k')\in\mathcal{I}} \mathcal{N}(w_{kk'};,\theta_{kk'},\sigma_{\mathbf{W}}^2). \tag{4.6}$$

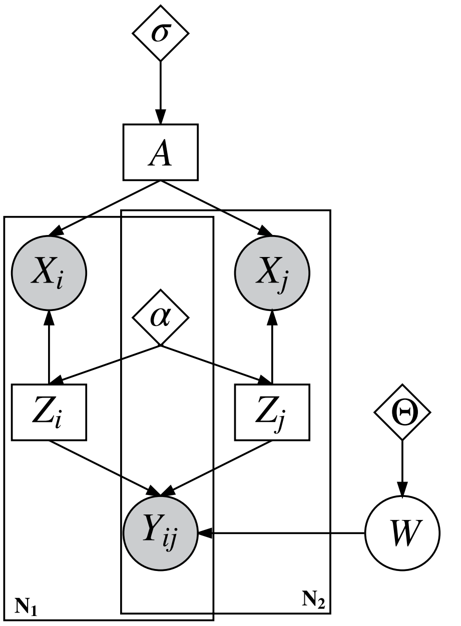


Figure 4.1 The plate notation of our model.

For simplicity, we let $\sigma_{\mathbf{W}}^2 = 1$. When both $z_{ik} = 1$ and $z_{jk'} = 1$, the element $w_{kk'}$ of $\mathbf{W}$ indicates the joint weight of the kth attribute in $\mathbf{z}_i$ and the k'th attribute in $\mathbf{z}_j$. Once $\mathbf{W}$ is determined, the prediction rule for our binary classifier becomes $\hat{y}_{ij} = \mathrm{sign}(\mathbf{z}_i \mathbf{W} \mathbf{z}_j^T)$. The putative pair-wise labels y^* for unknown pair-wise labels can also be generated with this prediction rule for Re-ID probability learning. Thus, the joint probability of the discriminative model becomes:

$$\mathrm{P}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = \prod_{(i,j)\in\mathcal{I}} \mathrm{P}(y_{ij}|\mathbf{Z},\mathbf{W}) \prod_{(i,j)\in\mathcal{U}} \mathrm{P}(y_{ij}^*|\mathbf{Z},\mathbf{W}). \tag{4.7}$$

In terms of representation learning, all the samples are combined and used for training, whether or not they have a known pair-wise label. Given this learning schema handles both labeled and unlabeled pairs, our algorithm can be considered for semi-supervised Re-ID tasks. Overall, our model is formulated as

$$\mathrm{P}(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{A},\mathbf{W}) = \mathrm{P}(\mathbf{X})\mathrm{P}(\mathbf{A})\mathrm{P}(\mathbf{Z})\mathrm{P}(\mathbf{W})\mathrm{P}(\mathbf{Y}|\mathbf{Z},\mathbf{W})\mathrm{P}(\mathbf{X}|\mathbf{Z},\mathbf{A}). \tag{4.8}$$

The related parameters have been omitted from for simplicity. In (4.8), both the representation learning and the re-identification learning models shared the same prior of latent factors $\mathrm{P}(\mathbf{Z})$. A plate notation of our model is illustrated in Figure 4.1.

4.2.2 Optimization

This section outlines the algorithms for efficiently learning the proposed Bayesian model in (4.8). The generative model for attribute learning is considered first. Using the priors from the last section, we have the joint distribution:

$$\frac{1}{(2\pi\sigma_{\mathbf{A}}^2)^{(K+D)/2}}\exp\{-\frac{1}{2\sigma_{\mathbf{A}}^2}\mathrm{Tr}(\mathbf{A}^T\mathbf{A})\}. \tag{4.9}$$

Following [6], we let $\sigma_{\mathbf{X}} \to 0$ and $\alpha = \exp(-\lambda^2/2\sigma^2\mathbf{X})$. Then

$$-\log \mathrm{P}(\mathbf{X},\mathbf{A},\mathbf{Z}) \sim \|\mathbf{X}-\mathbf{A}\mathbf{Z}\|_F^2 + \lambda^2 K_+, \tag{4.10}$$

where λ can be treated as a penalty parameter as K_+ increases. It is easy to verify that $\mathbf{A}$ has a closed formed solution when $\mathbf{Z}$ is fixed. Then, according to Bayesian theory, the posterior distribution of the uncertain remainder in (4.8) is

$$\mathrm{P}(\mathbf{Y},\mathbf{W}|\mathbf{Z}) = \mathrm{P}(\mathbf{Y}|\mathbf{Z},\mathbf{W})\mathrm{P}(\mathbf{W}). \tag{4.11}$$

According to the definition of re-identification possibility in (4.5) we have

$$-\log \mathrm{P}(\mathbf{Y},\mathbf{W}|\mathbf{Z}) \propto \sum_{(i,j)\in\mathcal{I}\cup\mathcal{U}}^{|\mathcal{I}|+|\mathcal{U}|} \mathrm{sign}(\mathbf{Z}_1\mathbf{W}\mathbf{Z}_2). \tag{4.12}$$

The remaining subproblem is to infer the probability $\mathrm{P}(\mathbf{W})$ when the other parameters are fixed. A straight forward method is to estimate a single value of $\mathbf{W}$ using $\mathrm{P}(\mathbf{W}) \propto \beta\|\mathbf{W}\|_F^2$ where β represents a leverage parameter as in previous works [22, 83]. However, our framework exploits the maximum entropy discrimination (MED) method [41] to learn the distribution of $\mathrm{P}(\mathbf{W})$. According to the MED theory, we can learn $\mathrm{P}(\mathbf{W})$ by estimating the expectation of $\mathbf{W}$ and solving the optimization problem

$$\min_{\mathrm{P}(\mathbf{W})\in\mathcal{P}} \mathrm{KL}(\mathrm{P}(\mathbf{W})||\mathrm{P}_0(\mathbf{W})) + C\mathcal{E}_\ell(\mathbb{E}(W)), \tag{4.13}$$

where $C > 0$ is a regularization parameter that leverages the influence of the prior and the max-margin hinge loss. $\mathcal{P}$ denotes the space of distributions of $\mathrm{P}(\mathbf{W})$. KL(p||q) denotes the Kullback–Leibler divergence, which is used to evaluate the distribution divergence between the distributions p and q. $\mathbb{E}(\mathbf{W})$ is the expectation of $\mathbf{W}$, and $\mathcal{E}(\cdot)$ is a loss function.

Algorithm 3: Semi-supervised Bayesian Attribute Learning

```
Initialize $K_+ = 1$,$\mathbf{A} = [\sum_i \mathbf{x}_i / N]$.
while objective value in (4.18) deceasing do
    for $n = 1, \cdots, N$ do
        for $n = 1, \cdots, K_+$ do
            Determine $z_{ij} \in \{0, 1\}$ to minimize the objective value in (4.18) greedily;
        end for
    end for
    $\mathbf{A} \leftarrow \mathbf{X}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}$.
    Sample a new basis $\mathbf{a}_{K_+}$ with probability
    $\mathrm{P}(\mathbf{a}_{K_+} = \mathbf{x}_i - \mathbf{A}\mathbf{z}_i) \propto \|\mathbf{x}_i - \mathbf{A}\mathbf{z}_i\|_2^2$.
    update $\mathbf{A} \leftarrow [\mathbf{A}, \mathbf{a}_{K_+}]$;
    update $K_+ \leftarrow K_+ + 1$.
    update $\Theta$ which is the expectation of $\mathbf{W}$ as in (4.17)
    update $\mathbf{y}^*$.
end while
```

Now, we turn to the error function. As a binary model, the training error of our model would be $\mathcal{E}_{tr} = \sum_{(i,j)\in\mathcal{I}\cup\mathcal{U}} \delta(y_{ij} \neq \hat{y}_{ij})$ where $\delta(\cdot)$ is an indicator function that equals 1 if the predicate holds, and 0 otherwise. However, the non-convexity of this error function makes it difficult to deal with, so instead we have used the well-studied convex hinge loss in our model as a surrogate loss $\mathcal{E}_\ell(\mathbb{E}(W)) = \sum_{(i,j)\in\mathcal{I}\cup\mathcal{U}} h_\ell(y_{ij} f(\mathbf{x}_i, \mathbf{x}_j))$, where $f(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i \mathbb{E}(\mathbf{W}) \mathbf{z}_j^T$ denotes the latent discriminant function [124]. After eliminating irrelevant terms, the subproblem can be written as

$$\begin{aligned} &\min_{\mathrm{P}(\mathbf{W})\in\mathcal{P}} \mathrm{KL}(\mathrm{P}(\mathbf{W})||\mathrm{P}_0(\mathbf{W})) + C \sum_{(i,j)\in\mathcal{I}} \xi_{ij} \\ &\forall (i,j) \in \mathcal{I}\cup\mathcal{U}, \text{ s.t. } y_{ij}(\mathrm{Tr}(\mathbb{E}(\mathbf{W})\mathbf{Z}_{ij}^*) \geq \ell - \xi_{ij}, \end{aligned} \tag{4.14}$$

where $\mathbf{Z}_{ij}^* = \mathbf{z}_j^T \mathbf{z}_i$ and $\{\xi_{ij}\}_{(i,j)\in\mathcal{I}\cup\mathcal{U}}$ are slack variables. According to Lagrangian duality theory, the optimal problem can be calculated by

$$\mathrm{P}(\mathbf{W}) \propto \mathrm{P}_0(\mathbf{W}) \exp\{ \sum_{(i,j\in\mathcal{I}\cup\mathcal{U})} \omega_{ij} y_{ij} \mathrm{Tr}(\mathbf{W}\mathbf{Z}_{ij}^*)\}, \tag{4.15}$$

where $\{\omega_{ij}\}_{(i,j)\in\mathcal{I}\cup\mathcal{U}}$. Let Θ be the expectation of $\mathbf{W}$, and the dual problem becomes

$$\begin{aligned}\max_{\omega} \ell \sum_{(i,j)\in\mathcal{I}} \omega_{ij} - \frac{1}{2}(\|\Theta\|_2^2) \\ \text{s.t.} \forall(i.j) \in \mathcal{I}\cup\mathcal{U},\ 0 \leq \omega_{ij} \leq C.\end{aligned} \tag{4.16}$$

This optimization problem can be solved by solving the equivalent primal problem

$$\begin{aligned}\min_{\Theta} \frac{1}{2}(\|\Theta\|_2^2) + C \sum_{(i,j)\in\mathcal{I}} \xi_{ij} \\ \forall(i,j) \in \mathcal{I}\cup\mathcal{U},\ \text{s.t. } y_{ij}(\text{Tr}(\Theta\mathbf{Z}_{ij}^*)) \geq \ell - \xi_{ij}.\end{aligned} \tag{4.17}$$

Eq. (4.17) can be efficiently solved as a standard binary SVM problem with a vectorized matrix $\mathbf{Z}$ and Θ [90] using public SVM solvers.[1] Once the optimal expectation of $\mathbf{W}$, i.e., Θ^*, has been derived and the distribution of $\mathbf{W}$ has been certified, $\mathbf{Z}$ can be updated by greedily minimizing the following joint objective loss function:

$$\|\mathbf{X} - \mathbf{AZ}\|_F^2 + \lambda^2 K_+ + \mathcal{E}_\ell(\Theta^*) + \frac{1}{2}\|\Theta\|_2^2, \tag{4.18}$$

where $\mathcal{E}_\ell(\mathbb{E}(W)) = \sum_{(i,j)\in\mathcal{I}\cup\mathcal{U}} h_\ell(y_{ij}(\mathbf{z}_i\Theta^*\mathbf{x}_j))$. As $K_+ \to \infty$, the algorithm alternately updates $\mathbf{A}$, Θ and $\mathbf{Z}$, along with the putative pair-wise labels y^*. The overall algorithm is provided as Algorithm. 3.

4.3 Experiments

4.3.1 Datasets

The following set of experiments compare the performance of various classical and state-of-the-art algorithms on two small-scale datasets and one large-scale dataset that are widely referred to in Re-ID studies. The **VIPeR** dataset [30] collects 1,264 images of 632 people from two non-overlapping camera views. There are two images of each person, each captured by a different camera. Variations in viewpoint and illumination conditions are frequent in VIPeR. We randomly select 316 people as the testing set for the experiment; the ramaining people were used as the training set. The **PRID** dataset [35] contains images of individuals from two distinct cameras. Camera B has

[1] For large-scale datasets, the numbers of their pair-wise labels are huge, we use a Stochastic Gradient SVM package *SvmSgd* : http://leon.bottou.org/projects/sgd.

captured 749 persons and Camera A records 385 persons. In the dataset, 200 peoples are captured by both cameras. We selected images of 100 people taken by both cameras as the testing sets for the experiment and used the remaining images for the training sat. The **DukeMTMC-reID** dataset [140] is a subset of the DukeMTMC dataset. It collects 1,404 Re-ID targets and 408 distractors. The dataset comprises 17,661 gallery images and 2,228 probe images captured by eight cameras , with 1404 individuals appearing in more than two cameras. We split the dataset equally using 702 people for the training set and 702 people for the testing set.

4.3.2 Evaluation Metrics and Preprocessing

We used a cumulative matching characteristic (CMC) curve and mean average precision (mAP) as performance evaluation metrics. Both are widely used in the evaluation of Re-ID models [138]. In mAP evaluation, average precision is calculated for each probe, and the mAP is then calculated across all probe images. CMC calculates the probability that an image in the first rank k gallery set matches the probe image. Unlike previous works, such as [50, 122] that rank gallery images according to their similarity with the probe image, our model ranks the gallery images according to their Re-ID probability $\mathrm{P}(y=1|\mathbf{Z}_{prob},\mathbf{W},\mathbf{Z}_{gallery})$. A higher probability implies the probe and the gallery image are more likely to be the same person.

In the experiments that test two-view Re-ID models, we randomly selected a set of images captured by one of the cameras to form the probe set. The images captured by the other camera view(s) were used as gallery images. Following the pre-processing procedure outlined in [66], all images were first rescaled to 224×224 pixels. Then, we extracted 2048 dimensional feature vectors from the images using a pre-trained ResNet-50 deep neural network [32]. We conducted experiments over ten splits and report the average results.

4.3.3 Supervised Person Re-ID

In the experiments, we first compare our algorithms with several supervised Re-ID models on the VIPeR and PRID datasets. As shown in Table 4.1, we compared SBAL with various metric learning Re-ID , metric learning Re-ID algorithms, and joint learning Re-ID methods. Some representative learning methods, such as LOMO [62], were included as feature generation methods in joint learning algorithms. Overall, we observed that most of the metric and representation learning Re-ID methods reported lower performance than the joint learning methods. Direct joint representation learning

Category	Dataset	VIPeR	PRID
metric learning for Re-ID	PRML[36]	27.0	4.8
	LMF[136]	29.1	12.5
	KISSME[51]	25.4	10.2
	kLFDA[122]	40.7	19.7
	KCCA[69]	37.2	14.5
	MLAPG[63]	40.7	16.6
Representation learning for Re-ID	DLLR[50]	38.9	25.2
	SSDAL[104]	37.9	20.1
Joint learning for Re-ID	LORAE[103]	42.3	18.0
	LOMO+KISSME[132]	34.81	-
	LOMO+kLFDA[132]	38.58	22.40
	LOMO+XQDA[62]	40.0	26.70
	LOMO+NullSpace[132]	42.28	29.80
	SSDAL+XQDA[104]	43.5	22.6
	ImprovedDeep[1]	34.81	-
	SBAL(Ours)	**45.2**	**32.4**

Table 4.1 Supervised Re-ID result of Rank One Matching Accuracy(%) on two benchmarks. Best result of each Re-ID algorithm is marked as bold numbers.

(e.g., LOMO) to metric learning Re-ID methods, i.e., KISSME [51] and kLFDA [122], enhanced the performance of metric learning Re-ID methods by 10% at most. In terms of attribute learning methods, deep attribute driven Re-ID (SSDAL) [104] and our algorithm delivered higher performance than the others. Moreover, our method consistently reported the best performance of all the algorithms on both the VIPeR and PRID datasets and surpassed SSDAL by at most 3%.

Category	Dataset	VIPeR	PRID
metric learning for Re-ID	RankSVM[91]	20.7	-
	KISSME[51]	18.5	5.1
	kLFDA[122]	27.5	14.1
	KCCA[69]	24.6	5.3
	MFA[122]	25.3	-
Representation learning for Re-ID	SSCDL[73]	25.6	-
	DLLR[50]	32.5	22.1
Joint learning for Re-ID	SBAL(Ours)	**33.6**	**24.4**

Table 4.2 Semi-supervised Re-ID results in terms of rank-1 matching accuracy(%) for VIPeR and PRID datasets. The best result from each Re-ID algorithm is shown in bold.

4.3.4 Semi-supervised Person Re-ID

In comparing our algorithms with other semi-supervised Re-ID models on the VIPeR and PRID datasets, we set two-thirds of the training data as unlabeled. As in previous works [50, 73], we also introduced supervised metric learning methods RankSVM, KISSME, kLFDA, KCCA and MFA as baselines. In the experiments, they are training with only the labeled data. we introduced the semi-supervised version of DLLR [50] as another baseline. As shown in Table 4.2, all semi-supervised methods demonstrated lower performance than supervised learning in Table 4.1. More specifically, supervised learning methods such as kLFDA, KCCA and KISSME We also observed that the representation learning Re-ID methods showed better performance than the metric learning methods. The reason for this could be that metric learning methods rely on pair-wise labels. Overall, our method consistently reported the best performance on both the VIPeR and PRID datasets, which implies that our algorithm is robust even with few labeled pairs.

Category	Dataset	mAP(%)	CMC R1 (%)
(1)	Attributes+KISSME[98]	12.83	21.97
	APR[66]	51.88	70.69
	ACRN[98]	**51.96**	**72.58***
(2)	BoW+KISSME[137]	12.17	25.13
	Basel.[138]	44.99	65.22
	LOMO+XQDA[62]	17.04	30.75
	SBAL(Ours)	**52.42***	**71.03**

Table 4.3 Attribute learning results on DukeMTMC-reID dataset. (1) Learning with predefined attributes (2) Learning with no pre-defined attributes. The best result for each category is in bold. The overall best results are marked with an asterisk (*)

4.3.5 Attributes Learning in Re-ID

We further compared our algorithms with several state-of-the-art attribute learning Re-ID methods on the large-scale dataset DukeMTMC-reID. We divide the comparison algorithms into two category, learning methods with pre-defined attributes and those without. The learning methods with pre-defined attributes included three algorithms. APR [66] utilizes manually annotated attributes from DukeMTMC-reID to enhance deep learning Re-ID . ACRN [98] trains an attribute classifier using separate Re-ID data from PETA [18], which is then used in the training stage to learn the attributes for DukeMTMC-reID and subsequently learn the Re-ID model. We also use attributes generated by ACRN as pre-defined attributes and combined them with KISSME as a baseline method, denoted as Attributes+KISSME. The learning methods without pre-defined attributes assume that no attribute information has been provided in the training stage. Following the settings in [137], we used BoW features and KISSME (BoW+KISSME) and LOMO features and XQDA (LOMO+XQDA) as the baseline methods for joint learning. We also included a recently presented method Basel.[138] as a baseline.

The mAP and rank one accuracy for CMC performance is listed in Table 4.3. our method delivered the best performance in the comparison between attribute learning methods without pre-defined attributes. Comparing the learning methods with pre-defined attributes, our method performed 2% worse than the state-of-the-art method, ACRN, in terms of rank-1 accuracy. It implies our algorithm is very comparable as our algorithm did not require any pre-defined attributes.

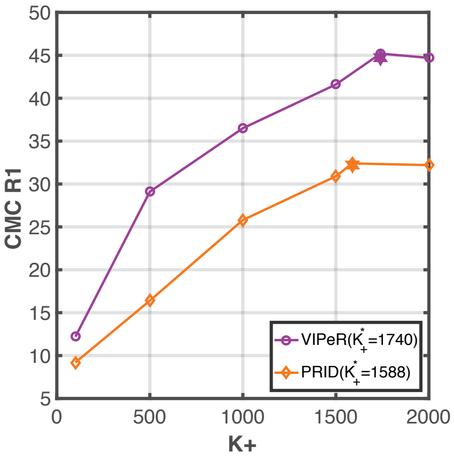


Figure 4.2 Influence of K_+ w.r.t. CMC Rank One accuracy. The automatically leaned attribute numbers are $K_+^* = 1740$ for VIPeR dataset and $K_+^* = 1588$ for PRID dataset(marked with asterisk symbol (*)).

4.3.6 The Influence of Latent Factor Dimensionality

To gauge the influence of automatically learned attributes, we used the settings specified for supervised learning on the VIPeR and PRID datasets and forced our algorithm to run after researching the optimal K_+ and stopped at $K_+ = 2000$. As Figure 4.2 shows, performance generally increased as K_+ increased. However, at an optimal $K_+^* = 1740$ for the VIPeR dataset and an optimal $K_+^* = 1588$ for PRID dataset, performance slightly degraded on both datasets. This implies that our algorithm is able to detect representative attributes with optimal numbers and can provide reliable Re-ID performance.

4.4 Summary

In this chapter, we proposed a novel semi-supervised Bayesian attribute learning framework, called SBAL, for person Re-ID . Through this framework, representation learning and Re-ID probability estimation are simultaneously optimized. The algorithm relies on semi-supervised learning to handle both labeled and unlabeled pairs of Re-ID data. It is based on factor-based attribute learning and can, therefore, adaptively learn binary latent factors that do not have pre-defined dimensionality. Through extensive experiments on two small datasets, we show that our algorithm outperforms various state-of-the-art methods. Further, the results reveal comparable performance on large-scale datasets without the pre-defined attribute information required by existing methods. For future works, we suggest extending our algorithm for non-linear applications by using deep generative models.

Chapter 5

Analogical Transfer learning

5.1 Background

Despite recent advances in machine learning applications, such as text and image classification, most conventional supervised learning algorithms can not offer satisfactory schema for learning new models from little data. Such challenge of learning from very few samples is refereed as few-shot learning or one-shot learning, and has attracted much attention in recent researches [11, 94, 102, 112, 119]. The main difficulty of few-shot learning is how to optimize the model when there comes new classes of data but few labeled training instances are provided for each class. Given sufficient pre-known labeled data from related domains, such problem can be addressed by transfer learning (TL) [88]. Transfer learning can benefit few-shot learning by transferring helpful prior knowledge from some pre-known source domains. With the prior knowledge from the source domains, the performance of the learning task in the target domain could be improved even with few samples [16, 88].

A critical problem in transfer learning for few shot learning is the negative transfer. A negative transfer is considered as an occurrence of the pernicious influence of performance when transfer knowledge from the source domain to the target domain [53, 88]. Previous TL methods attempt to relieve this problem mainly on the feature, instance or model/parameter level. In this chapter, we focus on the last two kinds of TL algorithms.

Instance transfer learning (ITL) assumes that negative transfers are caused by some of the source data that miss-leading the target task. Therefore, it suggests reweighing or selecting the source instances [88]. Typical ITL algorithms analyze the relativeness based on the representation or distribution between the source and the target domain. However, because there is few target instances, it is difficult to select source instances precisely. Moreover, the source data may not helpful or even may weaken the performance in practice, due to variously representation and distributions of real-world data [141].

Parameters transfer learning (PTL) (or model transfer learning) assumes that related tasks should have shared parameters or hyperparameters of a common prior distributions [88]. In the field, we refer a typical algorithms, Hypothesis Transfer Learning (HTL). HTL considers that the source *hypotheses*, e.g., classifiers, have been already well-trained and directly integrate them to the target hypothesis [53, 117]. The negative transfer in HTL is defined as a failure of satisfying the improvement condition (IC) [53], which assumes that the negative transfer occurs when hypothesis transfer cannot improve the performance of the target tasks. Since HTL treats the source hypotheses as already well-trained, it ignores the negative transfer caused by inconsistency of the source instances and the target instances, as analyzed in ITL [26]. Consequently, it cannot avoid to faced with the dilemma that even some of the source instances contribute to the source hypothesis, they may be harmful to the target hypothesis.

To relieve the aforementioned problem, in this chapter, we propose a novel algorithm, Analogical Transfer Learning (ATL). As in conventional transfer learning researches [88], we assume the target domain and the source domain are related but not the same. Therefore, only the source instances related the target instances can help the learning process in the target domain. We introduce an analogy strategy to transfer learning based on the cognitive theory of human beings [34, 110]. We propose the *analogy strategy* as a two-stage learning schema:

(1) **Revision**. First, the learner finds out the source instances according to their contributions directly to the target hypothesis. After that, the source hypotheses are learned with the selected source instances. In this stage, learner revises the source hypothesis (compared to the source hypothesis being learned from all of the source instances). The inconsistent knowledge to the target hypothesis is eliminated. Here, we borrow the term "revision" from the knowledge representation theory, in which such process is referred as *belief revision* [25]. In this stage, we relive the potential negative transfer on the instance level.

(2) **Transfer**. Second, the learner transfers the revised source hypothesis with the target hypothesis together to learn an *analogical hypothesis*. The analogical hypothesis is suitable for both the revised source hypotheses and the target hypothesis (trained with only the target instances). As a result, the source

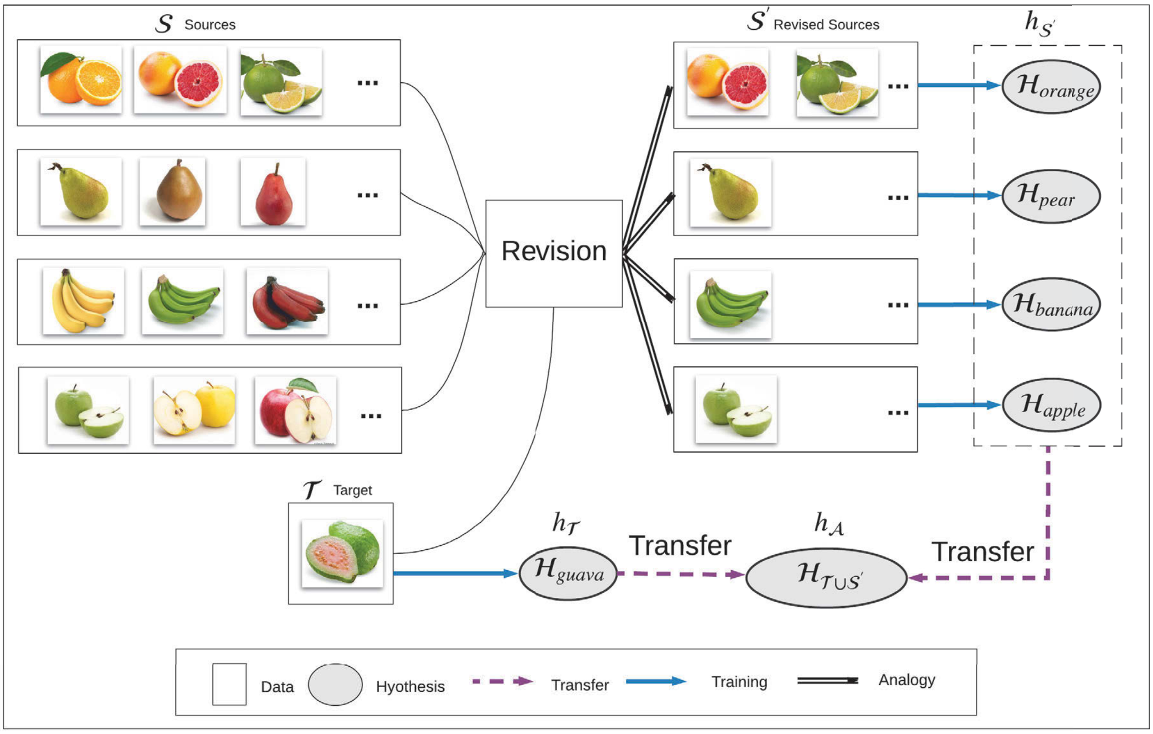


Figure 5.1 An example of our algorithm: 'Learning Guava from other fruits'. First, in the **Revision** stage, when a new (target) genus of fruit 'guava' comes, the learner selects the source instances related to the target instances (i.e., fruits related to guava) and learn the revised source hypotheses. Second, in the **Transfer** stage, an *Analogical Hypothesis* is concluded based on the observations of guava and the revised source hypotheses. Detailed discussion is in Chapter 2.1.

and the target hypothesis are consistent with the analogical hypothesis, and we, therefore, relive the potential negative transfer problem on the hypothesis level.

Generally, the proposed algorithm infers the properties of the target domain via learning from the source domain. In the algorithm, it propose to learn via a compassion of the structures and the features of the target and the source instances. Such that, we denote the proposed algorithm as an ‘analogy strategy’.

An example. We provide a illustration example of ‘Learning guava from other fruits’ via the analogy strategy in Fig. 5.1. First, in the Revision stage, when a new genus of fruit (target data) ‘guava’ comes, the learner first selects the source instances that are related to guava, such as ‘green apples’ and revises the source hypothesis ‘apple’ to ‘apples related to guava’. Second, in the Transfer stage, an analogical hypothesis is concluded based on the source hypothesis (e.g.,‘apples related to guava’) and the weak target hypothesis of "guava". Overall, on the target domain, the an analogical hypothesis is suit to the hypothesis of ‘guava’; on the source domain, it applies to the source hypothesis ‘apple related to guava.’

In the schema, we refer the final optimized hypothesis as the *analogical hypothesis*. The reasons are in two folds: First, different from previous works in HTL, our algorithm transfers knowledge from both the source hypothesis and the target hypothesis to optimize our hypothesis, rather than transfer the source hypothesis to the target hypothesis. Such that, the optimized hypothesis is different from the definition of ‘target hypothesis’ in HTL. Second, in the revision stage, useful knowledge contributes to the target hypothesis are picked out, which implies that they are analogically transferred to the target hypothesis. Notice that, rather than learning source hypothesis and directly use it to target task we learning a analogy hypothesis which is applicable for both source and target domain. The reason is that there only a few target data are provided and therefore the target hypothesis is very weak and biased.

5.2 Problem Statement

In this section, we first revisit the transfer learning schema then introduce our analogical transfer learning algorithm for few shot learning.

Notations. Let the superscript $^{\top}$ denote the transpose of a vector/matrix, $\mathbf{0}$ be a vector/matrix with all zeros, and $\|\mathbf{v}\|$ be the ℓ_2-norm of a vector $\mathbf{v}$. Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}^{\top} \in \mathbb{R}^{n \times d}$ be the data set with d features and n instances. Let $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}^{\top} \in \mathbb{R}^{n \times c}$ be the label matrix corresponding to $\mathbf{X}$ with c classes. More specifically, in transfer learning, we denote a *source domain data* as $\mathcal{S} = \{(\mathbf{x}_{S_1}, \mathbf{y}_{S_1}), \cdots, (\mathbf{x}_{S_n}, \mathbf{y}_{S_n})\}$ and a

Table 5.1 List of Important Notations and Descriptions

Notation	Descriptions	Notation	Descriptions
h	The hypothesis	$\mathcal{T},\mathcal{S}$	Target and Source domain data
$\mathbf{X}$	The instances	θ	Parameter of analogical hypothesis
$\mathbf{Y}$	The labels	β	The learning pace parameter
$\lvert\mathcal{A}\rvert$	Number of instances in $\mathcal{A}$	$\mathbf{v}$	The indicator vector of instances
$\mathbf{t}$	Parameters of $h_\mathcal{T}$	λ_j	Leverage parameter
$\mathbf{s}$	Parameters of $h_\mathcal{S}$	$\Phi(v)$	Regularization term to v

target domain data as $\mathcal{T} = \{(\mathbf{x}_{T_1}, \mathbf{y}_{T_1}), \cdots, (\mathbf{x}_{T_m}, \mathbf{y}_{T_m})\}$. We denote $h_\mathcal{S}(\cdot)$ and $h_\mathcal{T}(\cdot)$ are *hypotheses* on the source domain and the target domain respectively. Here a hypothesis can be a mapping from data space to label space, e.g. a classifier. In this chapter, we focus on the convex hypotheses, i.e. the hypotheses target to minimizes a convex combination of the empirical risks [53]. We denote $|\mathcal{A}|$ as the number of instances in $\mathcal{A}$. We denote $h_\mathcal{A} \mapsto h_\mathcal{B}$ a hypothesis transfer from $\mathcal{A}$ to $\mathcal{B}$. The symbol $\mapsto$ is a map from domain $\mathcal{A}$ to $\mathcal{B}$ on the hypothesis level. A list of important notations and descriptions is stated in Table 5.1.

Hypothesis Transfer Learning Revisit. Generally, a machine Learning task could be formed as an Empirical Risk Minimization (ERM) problem:

$$\mathbb{E}_\mathcal{T} = \min_\theta \sum_{(\mathbf{x}_i, y_i)\in\mathcal{T}} \mathcal{L}(h_\mathcal{T}(\mathbf{x}_i, \theta), y_i), \tag{5.1}$$

where $h_\mathcal{T}(\cdot)$ denotes a hypothesis leaned by domain data $\mathcal{T}$ and θ is the parameter. In normal machine learning task, $h_\mathcal{T}(\cdot)$ is learned and used in the same domain. However, there may be only a few data for training $h_\mathcal{T}$ such that the learning performance is hard to guarantee.

To relive this problem, one approach is transfer learning (TL). TL algorithms transfer knowledge from a source domain with sufficient knowledge to the target domain, in order to improve the performance of target task. Particularly, we could consider transfer knowledge on the hypothesis level by Hypothesis transfer learning (HTL) [53]. For simplicity, we first consider there is only one source domain, and will expend to the multi-source scenario in the next section. We aim to find a transfer of hypothesis from a source domain $\mathcal{S}$ to the target domain $\mathcal{T}$, i.e. $h_\mathcal{S} \mapsto h_\mathcal{T}$, to optimize the ERM of $h_{\mathcal{T}}$. Thus, the ERM problem becomes to:

$$\begin{aligned} \mathbb{E}_{\mathcal{T}} = \min_{\theta} & \sum_{(\mathbf{x}_i,y_i)\in\mathcal{T}} \mathcal{L}(h_{\mathcal{T}}(\mathbf{x}_i,\theta),y_i), \\ s.t. \;\; & h_{\mathcal{S}} \mapsto h_{\mathcal{T}}. \end{aligned} \tag{5.2}$$

HTL assumes that $h_{\mathcal{S}}$ is pre-trained and focus on pursuing an optimal transfer to improve the performance of $h_{\mathcal{S}}$ on parameter level. For the transfer of hypothesis, the previous works propose $h_{\mathcal{S}} \mapsto h_{\mathcal{T}}$ as $h_{\mathcal{T}}(\mathbf{x},\mathbf{u}) = \mathbf{u}^{\top}\mathbf{x} + h_{\mathcal{S}}(\mathbf{x},\mathbf{w})$. Theses works consider the hypotheses are all in form of a linear function. For non-linear tasks, kernel functions are used to project the data to high dimensional spaces.

5.3 The Proposed Method

5.3.1 Single-Source ATL

First, we start with the single-source ATL scenario. As discussed in former sections, HTL treats the source hypothesis as well-trained and the source instances are unaccessible. However, some of the source instances that contributes to the source hypothesis may harm the target hypothesis. Therefore, it is desired to eliminate the harmful samples when training the source hypotheses. To solve this problem, we suggest to learn a *analogical hypothesis* by a new transfer of hypothesis $\{h_{\mathcal{T}}, h_{\mathcal{S}'}\} \mapsto h_{\mathcal{A}}$, where $\mathcal{A} = \mathcal{S}^{'} \cup \mathcal{T}$ and $\mathcal{S}^{'} \subseteq \mathcal{S}$ is a revised subset of $\mathcal{S}$ that only helpful instances contributing to the target hypothesis are selected. The analogical hypothesis is not only applicable to the target hypothesis, but also applicable to revised source hypothesis. We state the ERM problem of our algorithm as:

$$\begin{aligned} \mathbb{E}_{\mathcal{A}} = \min_{\theta} & \sum_{(\mathbf{x}_i,y_i)\in\mathcal{A}} \mathcal{L}(h_{\mathcal{A}}(\mathbf{x}_i,\theta),y_i), \\ s.t. \;\; & \{h_{\mathcal{T}}, h_{\mathcal{S}'}\} \mapsto h_{\mathcal{A}}. \end{aligned} \tag{5.3}$$

where we denote θ as the parameter of $h_{\mathcal{A}}$. Next we will discuss the two-stage algorithm for ATL. As mentioned in former sections, the algorithms contains a revision stage for select source instances for revising source hypothesis and a transfer stage for learning the analogical hypothesis.

Revision. First, we learn an initialized target hypothesis utilizing all of the target instances. We determine the corresponding parameter $\mathbf{t}$ for $\mathcal{T}$ as follows:

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{T}} \mathcal{L}(h_{\mathcal{T}}(\mathbf{x}_i, \mathbf{t}), y_i). \tag{5.4}$$

where we denote $\mathbf{t}^*$ as the optimal parameter of $h_{\mathcal{T}}$. Eq. (5.4) can be solved as standard optimization problems depending on the base-learner hypothesis, e.g. SVM classifiers [33]. The parameter $\mathbf{t}^*$ will not be changed in the next stage. The base-learners can be standard SVM, kernel SVM or Least Squared SVM. The discussions of base-learners are in [53, 54].

In former works, source hypothesis is treated as unchangeable. It is trained by all the source instances. In our algorithm, we propose to revise this source hypothesis by selecting instance that contribute to the target hypothesis. To revised the source hypothesis, we introduce a self-paced learning (SPL) schema [52] for selecting $\mathcal{S}^{'} \subset \mathcal{S}$:

$$\min_{\mathbf{v}} \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{A}} v_i \mathcal{L}(h_{\mathcal{T}}(\mathbf{x}_i, \mathbf{t}), y_i) + \Phi(\mathbf{v}), \tag{5.5}$$

where v_i is the leverage parameter for each instance in $\mathcal{A} = \mathcal{S}^{'} \bigcup \mathcal{T}$ from a leverage vector $\mathbf{v} = \{v_1, \cdots, v_N\}$ and $\Phi(\mathbf{v})$ is the regularization term that controls the learning rate. For the purpose of analogy, we aim to analyze the contribution of only the source instances to the target hypothesis learning. Such that, we fix the indicator vector $v_i = 1$, $\forall (\mathbf{x}_i, y_i) \in \mathcal{T}$ to keep all target instances and only tune the indicator corresponding vector v_j, $\forall (\mathbf{x}_j, y_j) \in \mathcal{S}$. For simplicity, we denote $l_i = \mathcal{L}(h_{\mathcal{A}}(\mathbf{x}_i, \theta^*), y_i)$, let $|\mathcal{A}| = N$, and obtain:

$$\min_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} l_i v_i + \Phi(\mathbf{v}). \tag{5.6}$$

The classical regularization term in previous works [43, 44] are formed as hard/binary weighted to select examples. For instance, the previous regularization term in [44] is

$$\Phi(\mathbf{v}) = -\frac{1}{\beta} \sum_{j=1}^{N} v_j, \tag{5.7}$$

where $\beta \in (0, 1]$ is the learning pace parameter. In experiments β is gradually increased for studying from helpful to harmful source examples according to their contribution to target hypothesis.In SPL, it is also considered as “easy to complex” analysis. By setting the gradient with respect to v_j to zero in Eq. (5.6), the optimal weight for the

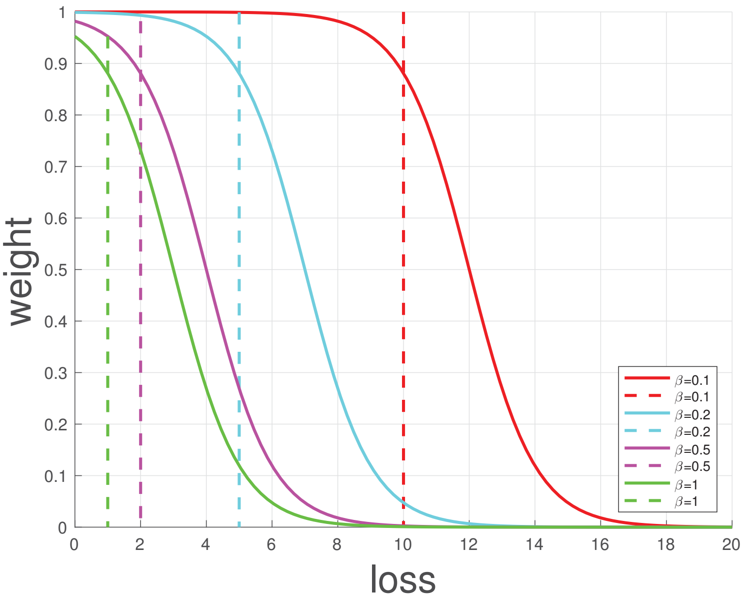


Figure 5.2 Parameter Sensitivity of self paced regularizer. Soft weight drops when loss values increasing (solid lines). When β becomes larger, the function becomes more sensitive, i.e. weights drop more dramatically even loss values are small. Hard weight (dash lines) do not sensitive while loss increases.

j-th example is

$$v_j^* = \begin{cases} 1 & if\ l_i \leq \dfrac{1}{\beta}; \\ 0 & if\ l_i > \dfrac{1}{\beta}. \end{cases} \tag{5.8}$$

However, in practice, since noise is usually non-homogeneously distributed in the data, it is hard to determine whether one example is easy or complex with hard weights. As shown in Fig. 5.2 (dash lines), when β becomes larger, when values of loss are small, they cannot be divided. In this case the self-paced scheme will be invalid.

In this chapter, we specialize the regularization term as a continuous function. The regularization form of Eq. (5.5) is

$$\Phi(\mathbf{v}) = -\sum_{j=1}^{N} f(v_j). \tag{5.9}$$

The regularization term of v_j is defined as

$$f(v_j) = v_j \ln(\frac{1-v_j}{v_j}) + \ln(\frac{1}{1-v_j}) + v_j(\frac{1}{\beta}+2). \tag{5.10}$$

By setting the gradient with respect to v_j to zero in Eq. (5.10), we obtain

$$v_j = \frac{1}{1+e^{l_i - \frac{1}{\beta} - 2}}. \tag{5.11}$$

where $\beta > 0$ is a sensitive controlling parameter. Soft weight can make our algorithm more sensitive when loss is small. As shown in Fig. 5.2 (solid lines), when β becomes larger, the weights are more sensitive to the loss value. Soft weighting is more effective than the hard weighting and can faithfully reflect the true importance of examples during training [43]. Moreover, by using a soft weighting, values of our function will be more distinct when β is small.

After obtained the weights of source instance, we can then select a subset of $\mathcal{S}' \subset \mathcal{S}$. The weights are directly related to the degree of the contribution of source instances to the target hypothesis. The revised source hypothesis can be learned:

$$\mathbf{s} = \arg\min_{\mathbf{s}} \frac{1}{|\mathcal{S}'|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}'} \mathcal{L}(h_{\mathcal{S}'}(\mathbf{x}_i, \mathbf{s}), y_i). \tag{5.12}$$

Transfer. In order to control the stability the transfer of hypothesis, we introduce a hypotheses transfer regularization term [53, 55] to Eq. (5.3):

$$\begin{aligned} &\min_{\theta} \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{A}} \mathcal{L}(h_{\mathcal{A}}(\mathbf{x}_i, \theta), y_i) + \Omega(\theta), \\ &s.t. \quad \{h_{\mathcal{S}'}, h_{\mathcal{T}}\} \mapsto h_{\mathcal{A}}. \end{aligned} \tag{5.13}$$

where we define the hypotheses transfer regularization term $\Omega(\theta) = \sum_{\mathbf{w}_j \in \{\mathbf{s},\mathbf{t}\}} \|\theta - \mathbf{w}_j\|^2$ to control the hypotheses dissimilarity between the pre-trained hypothesis $h_{\mathcal{S}'}$, $h_{\mathcal{T}}$ and the analogical hypothesis $h_{\mathcal{A}}$. Further, we formulate the transfer of hypothesis $\{h_{\mathcal{T}}, h_{\mathcal{S}'}\} \mapsto h_{\mathcal{A}}$ as a combination of an inner product and hypotheses $h_{\mathcal{S}'}$ and $h_{\mathcal{T}}$ as $h_{\mathcal{A}}(\mathbf{x}, \theta) = \theta^\top \mathbf{x} + h_{\mathcal{S}'}(\mathbf{x}, \mathbf{s}) + h_{\mathcal{T}}(\mathbf{x}, \mathbf{t})$, where θ is the parameter. In this equation, the parameter $\mathbf{s}$ and $\mathbf{t}$ are fixed after the revision stage. Further, for the loss function, we use ℓ_2-regularization as suggested in [117] for improvement of the generalization ability.

Overall, we formulate the optimization function of ATL as follows:

$$\begin{aligned} &\min_{\theta} \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{A}} \|h_{\mathcal{A}}(\mathbf{x}_i, \theta) - y_i\|^2 v_i + \sum_{\mathbf{w}_j \in \{\mathbf{s},\mathbf{t}\}} \lambda_j \|\theta - \mathbf{w}_j\|^2 + \Phi(\mathbf{v}), \\ &\qquad s.t. \quad h_{\mathcal{A}}(\mathbf{x}, \theta) = \theta^\top \mathbf{x} + h_{\mathcal{S}'}(\mathbf{x}, \mathbf{s}) + h_{\mathcal{T}}(\mathbf{x}, \mathbf{t}), \end{aligned} \tag{5.14}$$

where $\lambda_j > 0$ are leverage parameters. Since ℓ_2-norm is convex and all the considered hypothesis are either convex or linear combination of convex functions as we assumed, the proposed problem in Eq. (5.14) is solvable.

5.3.2 Multi-Source ATL

It is easy to expand our algorithm to multi-source domain scenario. A multi-source domain transfer learning is defined as that there are data from several different source domains in transfer learning. We denote the multi-source version of our algorithm as Multi-ATL (abbr. MtATL). For a k source domain scenario, we formulate the optimization problem for Multi-ATL as:

$$\begin{aligned} &\min_{\theta} \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{A}} \|h_{\mathcal{A}}(\mathbf{x}_i, \theta) - y_i\|^2 v_i + \sum_{\mathbf{w}_j \in \{\mathbf{S}, \mathbf{t}\}} \lambda_j \|\theta - \mathbf{w}_j\|^2 + \Phi(\mathbf{v}), \\ &\qquad s.t. \;\; h_{\mathcal{A}}(\mathbf{x}, \theta) = \theta^\top \mathbf{x} + \sum_{\mathbf{s} \in \mathbf{S}} h_{\mathcal{S}'}(\mathbf{x}, \mathbf{s}) + h_{\mathcal{T}}(\mathbf{x}, \mathbf{t}), \end{aligned} \tag{5.15}$$

where $\mathbf{S} = \{\mathbf{s}_1^{'}, \cdots, \mathbf{s'}_k\}$ is a set of parameters that contains all k source hypotheses and $\mathcal{A} = \{\bigcup_{i=1}^{k} \mathcal{S}_i^{'}\} \cup \mathcal{T}$. Each of $\mathbf{s}_k^{'} \in \mathbf{S}$ is determined separately in revision stage in the Multi-ATL algorithm. As already revised the source hypothesis for their best contribution to the target hypothesis in the revision stage, our algorithm does not require to determine a model selection parameter for each source hypothesis, which are hard to be weighted in former works [110].

Algorithm 4: Algorithm of Analogical Transfer Learning

```
Input: Source domain data $\mathcal{S}$, target domain data $\mathcal{T}$.
1: Initialize $\mathbf{v}$, $\beta$. Learn $h_{\mathcal{T}}$ by determining $\mathbf{t}$;
2: while not converge do
3:     (Revision) Update $\mathcal{S}^{'} \subset \mathcal{S}$ by determining $\mathbf{v}^*$, $\beta$ from solving subproblem (5.6) via
       Algorithm 5; Update $h_{\mathcal{S}'}$ by determining $\mathbf{s}$ on $\mathcal{S}^{'}$;
4:     (Transfer) Update $\theta^*$ by solving subproblem (5.14) according to Eq. (5.16);
5: end while
Output: Analogical hypothesis $h_{\mathcal{A}}$ with parameter $\theta^*$.
```

5.3.3 Optimization

In this section, we discuss the optimization method for the ATL problem following the two-stage learning algorithms.

Algorithm 5: Optimization Algorithm for optimal $\mathbf{v}^*$

```
Input: Training data $\mathbf{X}$, hypothesis parameter $\theta$ learning pace parameter $\beta$, threshold
    variable $\delta$.
1: Sort samples $\mathbf{x}_i \in \mathbf{X}$ in ascending order by their loss $l_i$. Accordingly, denote the labels
   and weights of $\mathbf{x}_i$ as $y_i$ and $v_i$.
2: for all $\mathbf{x}_i \in \mathbf{X}$ do
3:         Calculate score $v_i^*$ according to Eq. (5.11)
4:         if $v_i^* < \delta$ then $v_i = s_i$ else $v_i = 0$;
5: end for
Output: $\mathbf{v}^*$.
```

Revision. We start with the revision of the source hypothesis. First, as mentioned in last section, we first optimize $\mathbf{t}$ by utilizing all the target instances as in Eq. (5.4). Then, with $\mathbf{t}$ fixed, the optimization problem of Eq. (5.5) is simplified to Eq. (5.6), i.e., $\min_{\mathbf{v}} \frac{1}{N}\sum_{i=1}^{N} l_i v_i + \Phi(\mathbf{v})$. The regularization term of $\Phi(\mathbf{v}) = f(v_i)$ is derivable w.r.t v_i. Meanwhile, l_i is a convex loss function. Therefore, we could solve Eq. (5.6) by determining as mentioned in [43]. Particularly, for each sample $\mathbf{x}_i$, its corresponding optimal indicator value of $\mathbf{v}$ is determined as in Eq. (5.11). In practice, for computation efficiency we truncate v_i^* to zero by a selection control parameter δ. We illustrate the method in Algorithm 5. Finally, with $\mathbf{v}$ determined, we revise the source hypothesis $h_{\mathcal{S}'}$ by optimize its parameter $\mathbf{s}$ as formulated in Eq. (5.12).

Transfer. With $\mathbf{v}$ determined, the problem is simplified to Eq. (5.14). First, when $\mathbf{s}$ and $\mathbf{t}$ are determined, Eq. (5.14) has an closed form solution by calculating its derivative at θ:

$$\theta^* = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + N\sum\lambda_j\mathbf{I})^{-1}(\mathbf{y} - h_{\mathcal{S}'}(\mathbf{X},\mathbf{s}) - h_{\mathcal{T}}(\mathbf{X},\mathbf{t}) - \mathbf{X}^\top(\mathbf{s}+\mathbf{t})), \tag{5.16}$$

where θ^* is the optimal parameter of $h_{\mathcal{A}}$. Matrix $\mathbf{X}$ is a embedding matrix representing all training instances $\mathbf{x} \in \mathcal{A}$ and $\lambda_j > 0$ are leverage parameters for $h_{\mathcal{S}}$ and $h_{\mathcal{T}}$. Solving subproblem (5.14) is similar to conventional HTL. However, rather than transfer source hypothesis to target hypothesis, ATL transfers revised source hypothesis and target hypothesis to learn an analogical hypothesis. We illustrate the entire algorithm in Algorithm. 4.

5.4 Experiments

In this section, we conduct extensive experiments to validate the performance of the proposed algorithm on both synthetic and real-world datasets. On the synthetic dataset, we demonstrate the ability of the proposed algorithm to handle the negative transfer. On the real-world datasets, we use traditional SVM as a baseline. Note that it does not transfer any knowledge.

5.4.1 Baseline Algorithms

We compare the proposed algorithm with following transfer learning algorithms. We select algorithms that transfer knowledge on the instance, feature and parameter level. Joint transfer learning algorithms which attempt to joint transfer on multiple levels, e.g., feature and parameter, are also investigated. Classical classification algorithms SVM and TSVM are also compared in the experiments as baselines.

- **Baseline algorithms: SVM**: Support Vector Machine [33] is used as a baseline. We implement SVM using libsvm package [10] with a Gaussian kernel. In the experiments, it only uses source data for training and tests on the target domain. **TSVM** [45]: Transductive Support Vector Machine is a classical baseline of transfer learning.

- **Instance transfer learning algorithms: HATL** [46]: Hierarchical Active Transfer Learning exploits cluster structure shared by domains to perform transfer learning by leveraging source and a limited number of target domain data using active learning.

- **Feature transfer learning algorithms: KBTL** [28]: Kernelized Bayesian Transfer Learning finds a shared subspace of source and target domain by a kernel-based Bayesian dimensionality reduction model.

- **Parameter transfer learning algorithms: HTL** [53]: Hypothesis Transfer Learning transfer knowledge from source to target domain by optimizing target hypothesis as a combination of source domain and an inner product term. A model transfer regularization term is introduced to minimize the difference between the source and the target hypotheses. **PMT** [3]: Projective Model Transfer learning analyzes the angle between hyperplanes of source and target domain hypotheses and adopts it as a regularization term to standard SVM. **GreedyTL (abbr. GdTL)** [54]: It is a multi-source hypothesis transfer learning method with non-negative smooth loss function and convex regularization term. **Multi-KT (abbr. MtKT)** [110]: Multi-model Knowledge Transfer algorithm uses a least square SVM adoptive method to operate a multi-source transfer learning that can handle few training examples from multiple models.

- **Joint transfer learning algorithms: MMDT** [37]: The Max-Margin Domain Transforms algorithm jointly optimizes over a feature transformation mapping target domain data and classifier weights to the source domain.

5.4.2 Datasets and Setting

We first conduct an evaluation experiment on synthetic data. Then we run comparison experiments on real-world data.
1). Synthetic Data. Following the experiment in [7], synthetic source and target data are independently drawn from a double-moon distribution by 1000 random sampled instances. The source data is a distribution rotated 60° in a counterclockwise direction from the target domain distribution. Due to the rotation, the source and target domain results to exhibit different distributions. Indeed, the greater the rotation angle, the larger the discrepancy of the domains [7]. Labeled training instances are sampled only 20 in each experiment. They are marked as magenta points in Fig. 5.3(a).
2). Text Data. We perform our algorithms on two real-world text data sets. **(i) Newsgroup**: The Newsgroup [57] dataset contains approximately 20,000 newsgroup documents partitioned across 20 different subtopics of newsgroups, around 2,000 documents for each. For several subtopics, there is a high-level category, such as sci.crypt and sci.electronics are all belong to category Sci (S), etc. [24]. As in [24], we divide the data by their subtopics from three categories,i.e., Comp (C), Sci (S), Rec (R), and Talk (T). The details of categories are presented in Table 5.2. **(ii) Reuters**: The Reuters-21758 corpus dataset [59] contains news articles of Reuters website from three categories, Orgs (O), People (Pe), Place (Pl). Each category contains a total number of around 1,200 documents from different subtopics. Similar to on Newsgroup, single source tests and multi-source tests are run as three binary classification tasks aiming to divide different categories. Specifically, the classification tasks are run on 'O vs. Pe', 'O vs. Pl' and 'Pe vs. Pl'. The source and the target data are drawn from different subtopics under the same categories. We only sample 10 labeled training samples from target domain data randomly without replacement. Details are omitted as in [24] as there are much more subtopics.

Settings: For both of the text datasets, we perform single-source and multi-source test. **(i) Single-source Test:** First, we run a single-source test. Binary classification tasks are then performed aiming to classify different categories. **(ii) Multi-source Test:** We conduct multi-source test for the purpose of exploring the performance of multi-source transfer learning. In the test, the source and the target samples are still drawn from the different subtopics under the same categories. Different from

Table 5.2 Data Description of Newsgroup

Task	Source	Target
C vs S	comp.graphics comp.os.ms-windows.misc sci.crypt sci.electronics	comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x sci.med sci.space
R vs T	rec.autos rec.motorcycles talk.politics.guns talk.politics.misc	rec.sport.baseball rec.sport.hockey talk.politics.mideast talk.religion.misc
R vs S	rec.autos rec.sport.baseball sci.med sci.space	rec.motorcycles rec.sport.hockey sci.crypt sci.electronics
S vs T	sci.electronics sci.med talk.politics.misc talk.religion.misc	sci.crypt sci.space talk.politics.guns talk.politics.mideast
C vs R	comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware rec.motorcycles rec.sport.hockey	comp.os.ms-windows.misc comp.windows.x rec.autos rec.sport.baseball
C vs T	comp.graphics comp.sys.mac.hardware comp.windows.x talk.politics.mideast talk.religion.misc	comp.os.ms-windows.misc comp.sys.ibm.pc.hardware talk.politics.guns talk.politics.misc

the single-source test, the multi-source hypothesis is learned on multiple subtopics. For example, in Newsgroup on ‘C vs. S’, comp.graphics and sci.crypt are treated as positive and negative samples for one source data while comp.os.ms-windows.misc and sci.electronics are treated as another source data. In the experiments, we only sample 10 labeled training samples from the target domain data randomly without replacement for both the single-source test and the multi-source test.

3). Image Data. We perform our algorithms on one real-world image data set. **AwA:** Animals with Attributes (AwA) data set contains 30,475 images of 40 subclasses of different animals. For each sample, 4,096 features are learned and extracted by a VGG-19 deep neural network [101]. We collect six groups of animals as positive samples according to their biological families, i.e., Cetacea (whale-like), Amphibian (beavers, etc.), Felidae (cat-like), Canidae (dog-like), Muridae (mouse-like). On average, each group contains approximately 2,000 images of 5 subclasses of animals. The rest images of animals are used as negative samples. Similar to the experiment on text data, we produce the source data and the target data by subclasses of animals.

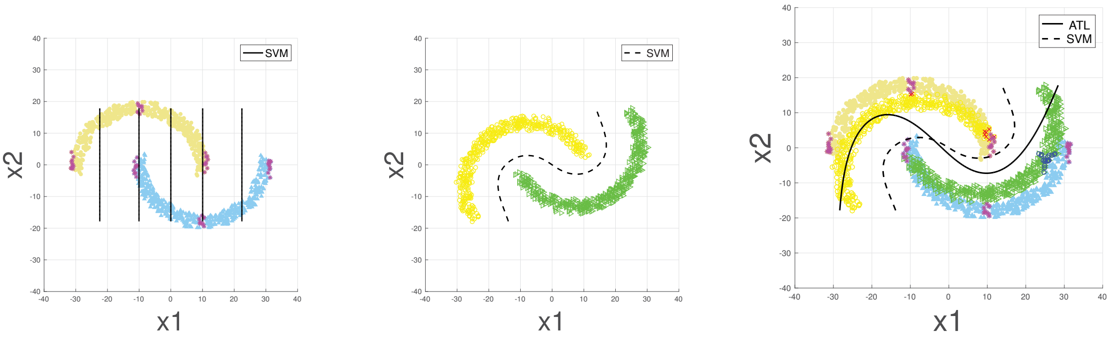
Settings: We perform inner-family and cross-family test on image data. **(i) Inner-family Test.** In the inner-family test, source and target samples are selected as what we did in text data experiments. For each task, the source and the target samples are drawn from different subclasses under the same families. **(ii) Cross-family Test.**To test the abilities of generalization, we additionally operate a cross-family test on image data. For each task, source and target samples are drawn from different families. It brings a challenge that source and target domains are much more different than the inner-family test. As a result, samples contribute to a well-trained source hypothesis may not benefit learning of analogical hypothesis, i.e., it would bring more negative transfers. In each experiment, we alter the number of positive target training data in a range of $\{1, 5, 10, 15\}$ with the equal number of negative samples. The performances are reported in terms of average results.

5.4.3 Performance Evaluation

1). Results on Synthetic Data

The performance of classification of our algorithm on synthetic data is shown in Fig. 5.3.

(1) Our algorithm can enhance the performance on the target domain with few training samples. As shown in Fig. 5.3(a), in the target domain, when labeled target training data is insufficient ($|\mathcal{T}| = 20$, marked as magenta points), hypothesis learned only on these samples fails to divide target samples correctly (boundary



(a) Target data (b) Source data (c) Analogical Transfer learning

Figure 5.3 Performances on synthetic data. ***Left***: In the target domain. Given only a few target samples (magenta points), the standard SVM classifier fails to classify the two half circles distribution (the decision boundary is marked as solid lines). ***Middle***: In the source domain. Given sufficient source samples (yellow and green points), the standard SVM classifier can classify two half circles distribution (The decision boundary is marked as dash line). ***Right***: If use entire source samples as in the Middle figure, it will fall to divide target examples that near to each other as shown in the center of graph. (The decision boundary is marked as dash line). Given selected source samples (dark blue points) and few target samples (magenta points), in our algorithm, the classifier can classify two half circles by a decision boundary (The decision boundary is marked as solid line).

marked as the solid line). Meanwhile, when labeled training data is sufficient in the source domain, it is easy to divide two class with source hypothesis (a kernel SVM base-learner in our experiments) in Fig. 5.3(b). Finally, as shown in Fig. 5.3(c), our algorithm can solve the classification problem by learning an analogical hypothesis (boundary marked as the solid line) even with only few examples ($|\mathcal{T}| = 20$, marked as magenta points).

(2) We also display that negative transfers are relieved by revision of source hypothesis. First, use all samples to train the source hypothesis and use it in the target domain will bring negative transfers. As shown in Fig. 5.3(c), the decision boundary of source hypothesis (dash line) fails to divide the two classes, even source data could be successfully classified in the source domain in Fig. 5.3(b). On the other hand, in our algorithm, only the source samples that contributes to the target hypothesis are picked up to revise the source hypothesis. Then, an analogical hypothesis is learned not only improved the performance of the target hypothesis but also the source hypothesis. It is shown in Fig. 5.3(c), the decision

Table 5.3 Performance (ACC ± standard deviation) (%) of single-source test

Dataset	Newsgroup						Reuters		
Task	C vs S	R vs T	R vs S	S vs T	C vs R	C vs T	O vs Pe	O vs Pl	Pe vs Pl
SVM	62.82±0.1	60.56±0.2	61.34±0.2	60.12±0.3	61.94±0.2	61.81±0.2	61.79±0.2	60.84±0.2	61.66±0.1
TSVM	66.52±0.1	66.21±0.3	67.85±0.2	66.00±0.3	65.99±0.2	69.17±0.2	68.57±0.1	63.66±0.1	62.26±0.1
HATL	66.90±0.2	64.02±0.1	62.54±0.3	63.97±0.3	69.12±0.2	64.87±0.2	88.92±0.2	88.87±0.1	89.06±0.2
KBTL	65.83±0.3	67.43±0.2	62.77±0.2	67.86±0.1	69.02±0.2	63.00±0.2	79.47±0.2	74.44±0.2	75.34±0.1
PMT	63.73±0.1	60.25±0.1	64.25±0.3	61.33±0.3	65.17±0.2	56.65±0.2	67.96±0.1	67.68±0.1	62.26±0.2
HTL	67.81±0.2	66.69±0.1	65.70±0.3	66.56±0.3	66.71±0.2	66.58±0.2	75.81±0.2	74.51±0.2	74.94±0.1
ATL	**67.99±0.2**	**67.96±0.1**	**69.07±0.2**	**67.67±0.2**	**69.66±0.1**	**67.14±0.2**	**77.20±0.1**	**76.01±0.1**	**78.00±0.2**

boundary of the analogical hypothesis (solid line) successfully divides the two classes.

2). Results on Text Data

Accuracy (ACC) is chosen as the evaluation measurement of the performances. Each experiment is called ten times independently, and we report the average results with corresponding standard deviations. In the experiments, we observe that:

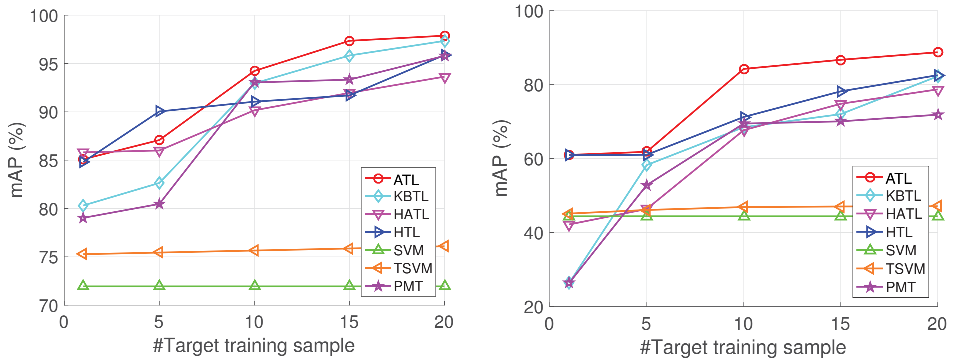
(1) Overall, all transfer learning algorithms shows better performances than baseline SVM, which runs with no transfer of knowledge. It implies that transfer knowledge from the source domain can help improve classification task on the target domain.

(2) In single source test, as presented in Table 5.3, our algorithm (ATL) outperforms all the compared algorithms. More specifically, our algorithm (ATL) consistently shows around 3% better performances than its counterparts on both Newsgroup and Reuters datasets. It implies that our algorithm is better in generalization in single source transfer learning scenario. Meanwhile, our algorithm consistently outperforms the HTL algorithm, which uses all source samples, on both Newsgroup and Reuters datasets. It implies that our algorithm can control negative transfer through self-paced sample selection schema.

(3) In the multi-source test, as presented in Table 5.4, our algorithm (ATL) outperforms all the compared algorithms give multiple source domains. More specifically, hypothesis transfer algorithms GreedyTL and Multi-ATL report a better performance in all algorithms. However, our algorithm (Multi-ATL) shows a better performance than its counterparts. It implies that our algorithm could control negative transfer and is better in generalization in multi-source transfer learning scenario.

Table 5.4 Performance (ACC± standard deviation) (%) of multi-source test

Dataset	Newsgroup						Reuters		
Task	R vs T	C vs S	C vs T	C vs R	S vs T	R vs S	Pl vs Pe	Pe vs Pl	O vs Pl
SVM	62.93±0.1	63.15±0.1	61.89±0.2	62.81±0.2	61.84±0.2	61.81±0.2	67.86±0.2	63.27±0.2	61.42±0.1
MMDT	63.08±0.2	63.87±0.1	65.18±0.0	64.36±0.1	65.81±0.0	63.91±0.3	67.33±0.1	69.74±0.2	63.87±0.1
MtKT	64.93±0.2	65.08±0.1	65.23±0.3	63.97±0.3	64.71±0.2	64.87±0.2	73.14±0.3	72.36±0.2	73.87±0.1
GdTL	67.26±0.2	63.55±0.2	73.57±0.2	71.46±0.2	65.67±0.2	63.35±0.1	78.98±0.1	75.57±0.4	80.05±0.2
MtATL	**70.85±0.1**	**70.36±0.1**	**77.67±0.1**	**73.92±0.2**	**71.54±0.1**	**73.14±0.1**	**81.79±0.1**	**80.74±0.1**	**81.35±0.1**



(a) Inner-family test (b) Cross-family test

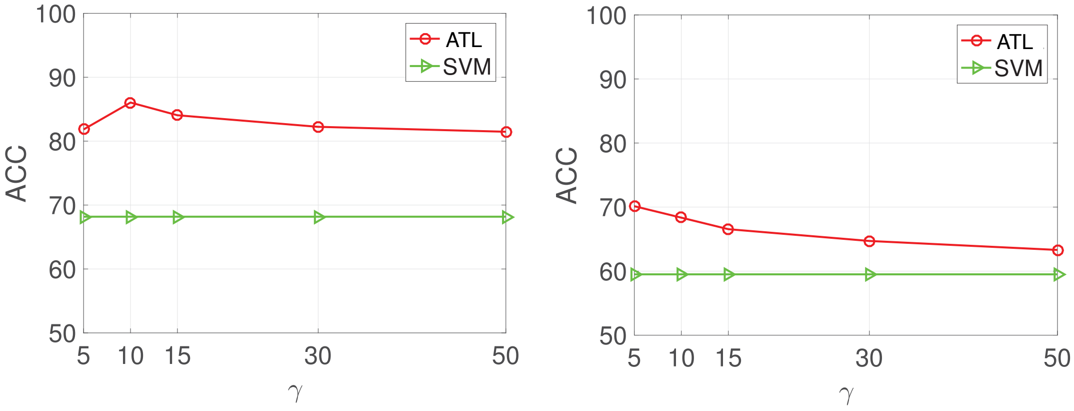
Figure 5.4 Performance on Image Data (1) Baseline algorithms SVM is not affected much by the number of target training samples; (2) Overall, all algorithms perform better in Inner-family test than Cross-family test; (3) Algorithms that use target domain knowledge generally perform better when given more target training samples; (4) Our algorithm outperform others early when given over 10 target training samples.

3). Results on Image Data

Mean Average Precision (mAP) is used as evaluation measurement of the performances. Each experiment is called ten times independently, and we report the average results with corresponding standard deviations. As Shown in Fig. 5.4, we can observe that:

(1) Overall, all experiments show that the performance is improved when the number of target training data is increasing. It is reasonable as target domain information contributes to the training of analogical hypothesis. Meanwhile, inner-family test results are better than cross-family results. This is not surprising, as different subclasses of animals in the same family may have similar features. On the other hand, since animals from different family may not have similar features, the

hypothesis trained with cross-family source will not perform as well as trained in inner-family source.



(a) γ w.r.t. ACC(%) on Reuters (b) γ w.r.t. ACC(%) on Newsgroup

Figure 5.5 Parameter Sensitivity of learning rate γ on Text Data. Our algorithm ATL (in red lines) performs consistently better than baseline SVM (green lines) at about 10% in Accuracy. Global optimal γ differs on different data sets.

(2) Our algorithms (ATL) (solid red line) performs better when given more than only ten target training samples. Moreover, it consistently performs better than HTL algorithm around 5% when given more than ten target training samples. It implies that our algorithm can control negative transfer through self-paced sample selection schema and improve hypothesis transfer learning.

(3) Our algorithms (ATL) (solid red line) not only is the best in the inner-family test but also significantly outperforms others in the cross-family test. It implies our algorithm is better in generalization.

4). Parameter Sensitivity of Learning Rate

We also test the influence of learning rate on text datasets. The number of target training samples is fixed at 10. As shown in Fig. 5.5, the learning rate γ is tuned in region $\{5, 10, 15, 30, 50\}$. We observe that on Reuters the corresponding accuracies are { 81.83, 86.04, 84.05, 82.24, 81.47} in which the peak result is reported when $\beta = 10$. On Newsgroup, it outputs { 70.14, 68.35, 66.55, 64.71,6 3.31} in which $\beta = 5$ leads the performances. Overall, a small γ seems better than larger β. However, the optimal choice is varied on different datasets. Overall, our algorithm is consistently outperformed better than the baseline algorithms.

5.5 Summary

In this chapter, in order to solve the problem of learning with few-shot text and image classification, we have proposed a novel analogical transfer learning algorithm. Rather than transferring knowledge from the source hypothesis to learn the target hypothesis, ATL learns an analogical hypothesis from both source and target hypothesis. Also, ATL was able to revise the source hypotheses by select helpful source instances according to their contribution to the target hypothesis. As a result, the proposed algorithm efficiently controls the occurrence of the negative transfer on both instance and hypothesis level. Extensive Experiments on synthetic and real-world datasets presented a consistent and reliable performance. Moreover, ATL can be easily expanded to the multi-source scenario.

For the future work, we suggest to investigate how to expand our algorithm by integrating non-linear hypotheses; Another attractive direction is to theoretically analyze the stability of our algorithm. For other machine learning applications, given the flexibility of our algorithm, we suggest to investigate the transfer between more challenging domains such as from image to video space, etc.

Chapter 6

Domain-aware Unsupervised Cross-dataset Transfer Learning

6.1 Background

Most existing Re-ID methods are supervised algorithms. They require labeling pairwise images across camera views for training Re-ID model. However, manually labeled Re-ID data is hard to produce. On the one hand, it is a tough task even for the human to annotate the same person in different camera views among a huge number of imposters [49, 89]. Meanwhile, camera numbers are increasingly large in today's world, e.g. over a hundred in an underground station. It makes the labeling cost becoming prohibitively high because supervised Re-ID methods require sufficient label information for each pair of camera views. As a result, the scalability of supervised Re-ID methods is severely limited and hard to applied to practical Re-ID applications.

To overcome the limitation of supervised Re-ID methods, one solution is to perform the identification with unsupervised learning algorithms, which utilizes only unlabeled data. However, typical unsupervised methods often are proposed for a single dataset. Without labels for matching information, unsupervised Re-ID methods sometimes are unable to recognize persons across camera views because of the uncontrollable and/or unpredictable variation of appearance changes across camera views, such as body pose, view angle, occlusion and illumination conditions [47, 89, 120], etc. As a result, most of the single-dataset unsupervised Re-ID methods report significantly worse performance than supervised methods.

Few recent works are proposed to address unsupervised person re-identification problem via cross-dataset transfer learning methods [77, 78, 89]. They intend to capture dataset-invariant and discriminative representations across multiple datasets. Different

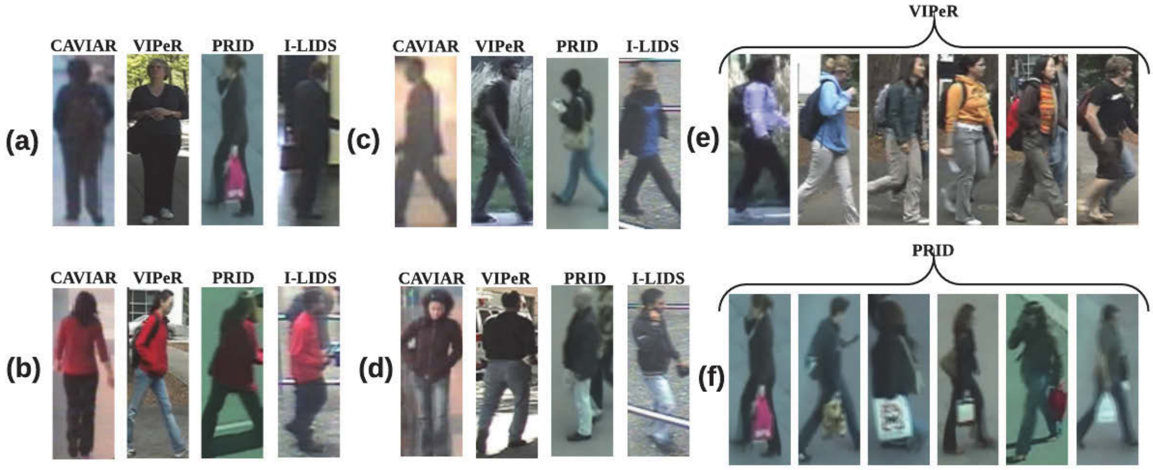


Figure 6.1 Examples of common and unique appearances on four datasets. Better being watched in color. **Common appearances** shared by four datasets: (a) wearing dark coat and pants; (b) wearing red upper cloth; (c) walking forward; (d) wearing dark cloth and light color pants. The images in (a) to (d) from different datasets and they are belong to the different persons. **Unique appearances**: (e)VIPeR: carrying backpack; (f)PRID: carrying bags in hand.

from single-dataset works, cross-dataset transfer learning brings an incredible challenge in Re-ID . First, it requires completely different learning task under different domains, i.e. identifying sets of non-overlapped persons under different camera networks. Second, they are also required to learning discriminative presentations on the target dataset, which may be heavily affected by the source datasets. In the research, we observe that among the Re-ID datasets there are not only shared common appearances [89, 120] but also domain-unique appearances.

As presented in Fig. 6.1, we illustrate some instances of Re-ID datasets VIPeR[30], PRID[35], iLIDS[139] and CAVIAR[14]. First, as shown in Fig. 6.1, (a) to (d), there are common appearances across domains, such as 'wearing black cloth' or similar pose such as 'walking forward'. Second, domain-unique appearances also be observed. As shown in Fig. 6.1, (e) and (f), many individuals are carrying backpacks in VIPeR dataset while many persons are captured carrying a handbag in PRID dataset. The reason is that cameras are set up the different scenes, such as shopping mall, campus, and airport. Therefore, simply ignoring or disregarding the importances of domain-unique features definitely will degraded the Re-ID performance. As we will show in the experiments, previous algorithms relying on the common appearance will mismatch the lady to other person wear dark cloth even that they are clearly not carrying a pink handbags.

However, previous works do not value the importances of domain-unique appearances [89].

The best attempt is a recent work in [89], which proposes an unsupervised cross-dataset dictionary learning methods. The algorithm learns a common dictionary across multiple datasets. With the dictionary, it can encode samples of Re-ID observations in a low dimensional space. Additionally, it also learns unique dictionaries for each dataset but separates them with the shared dictionary. As a result, it fails to leverage the importances of the common and the domain-unique representations. As we will show in the experiments, such algorithm will rely on common appearances rather than on domain-unique appearances, even such appearances are obviously distinct between persons.

To overcome the mentioned limitations and improve the Re-ID performance, we propose our algorithm. It is an domain-aware unsupervised cross-dataset multi-task learning algorithm. Our algorithm not only can obtain shared appearances across datasets via multi-task dictionary learning but also captures the domain-unique appearances. Rather than using Euclidean distances, we bring discriminative overlapping support as the metric of inter-dataset similarity. The importance of common and domain-unique appearances are valued simultaneously and jointly contribute to the representation learning for Re-ID task. We illustrate the procedure of our algorithm in Fig. 6.2. The main contributions of our algorithm are stated in three folds as following:

- We propose a novel unsupervised cross-dataset learning algorithm with support discriminative regularization for person Re-ID . To our knowledge, it is the first attempt to leverage the common and domain-unique representations across datasets in the unsupervised Re-ID application.
- We introduce an iterative re-weights optimization scheme to solve our problem. Our algorithm simultaneously optimizes the common representation and minimizes the overlapping supports across datasets to enrich the domain-unique representations.
- Extensive experimental studies on benchmark datasets show superior performances of our algorithm over state-of-the-art algorithms.

6.2 The Proposed Framework

6.2.1 Formulation

We focus on the cross-dataset person re-identification problem that the datasets are collected from several different camera networks. In multi-task learning, we are able to

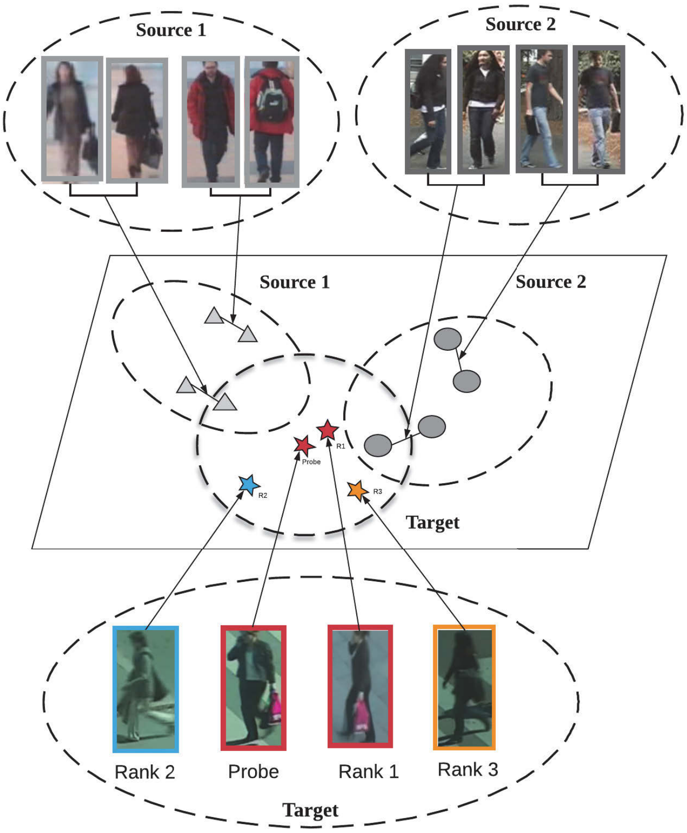


Figure 6.2 Overview of our algorithm for person Re-ID . Source datasets are datasets with labels, target dataset is a dataset with no label. A dictionary is learned with all datasets. With the dictionary, all images are encoded in one low dimensional space. The linked samples in the source dataset are labeled as the same person.

learn and transfer knowledge of the labeled source datasets to the unlabeled target dataset and therefore overcome the limitations of unsupervised Re-ID problem. Such kind of algorithm is claimed as an unsupervised [78, 89] and asymmetric [133] multi-task leaning algorithm.

Let $\mathbf{X} = \{\mathbf{X}_{S_1}, \cdots, \mathbf{X}_{S_N}, \mathbf{X}_T\} \in \mathbb{R}^{d\times\{n_1,\cdots,n_N,n_T\}}$ be the $T = N+1$ datasets with N *source* datasets and one *target* dataset. Each datasets has d features and n_t instances. We denote the dictionary shared by all datasets as $\mathbf{D} \in \mathbb{R}^{d\times k}$. With the dictionary $\mathbf{D}$, every image with feature vector $\mathbf{x}_{t,i}$, i.e. person appearance in original datasets, can be encoded as a sparse column atom in the coefficient matrix $\mathbf{A} = \{\mathbf{A}_{S_1}, \cdots, \mathbf{A}_{S_N}, \mathbf{A}_T\} \in \mathbb{R}^{k\times\{n_1,\cdots,n_N,n_T\}}$ in one lower k-dimensional subspace. Notice these corresponded representations are invariant to the camera view changes or camera network differences, makes it suitable for person re-identification.

Generally, we formulate the person re-identification task as the following multi-task dictionary learning problem:

$$\min_{\mathbf{D},\mathbf{A}} \sum_{t=1}^{T} R(\mathbf{X}_t, \mathbf{D}, \mathbf{A}_t) + \Omega(\mathbf{A}), \tag{6.1}$$

where T is the number of tasks. $\mathbf{D}$ is a dictionary shared by all tasks. We denote the reconstruction loss $R(\mathbf{X}_t, \mathbf{D}, \mathbf{A}_t)$ as the Frobenius norm:

$$R(\mathbf{X}_t, \mathbf{D}, \mathbf{A}_t) = \|\mathbf{X}_t - \mathbf{D}\mathbf{A}_t\|_F^2. \tag{6.2}$$

We denote $\Omega(\mathbf{A})$ as is the regularization term of $\mathbf{A}$. In our work, we specified the regularization term of our algorithm as a combination function of three regularization terms for person re-identification:

$$\Omega(\mathbf{A}) = \alpha g(\mathbf{A}) + \beta l(\mathbf{A}) + \gamma f(\mathbf{A}), \tag{6.3}$$

where α, β,γ are leverage parameters.

Structure sparsity. For the first regularization term, we bring in the structure sparsity by defining

$$g(\mathbf{A}) = \sum_{t=1}^{T} \|\mathbf{A}_t\|_{2,1} = \sum_{t=1}^{T} \sum_{k=1}^{n_t} \|\mathbf{a}^{(tk)}\|_2, \tag{6.4}$$

where $\mathbf{A}_t$ is the coefficient matrix for the t-th task and $\mathbf{a}^{(tk)}$ denotes the k-th row of matrix $\mathbf{A}_t$. We use a $\ell_{2,1}$-norm rather than ℓ_1 norm in typical dictionary works [31, 50] to gain row sparse representations. Such that, the proposed algorithm can find the shared nonzero supports shared all tasks automatically [74]. Moreover, as discussed in [139], $\ell_{2,1}$-norm can enhance the robustness and suppresses the affect of outliers. Outliers are widely appeared in Re-ID [49].

Pairwise relationship preserving. The second regularization term is defined as a Graph Laplacian term:

$$l(\mathbf{A}) = \sum_{t=1}^{T} \sum_{i,j=1}^{n_t} w_t(i,j) \| a_{t,i} - a_{t,j} \|^2 = Tr(\mathbf{A}_t \mathbf{L}_t \mathbf{A}_t^{\mathsf{T}}). \tag{6.5}$$

where $\mathbf{a}_{t,i}$ and $\mathbf{a}_{t,j}$ are column atoms of $\mathbf{A}_t$ and $\mathbf{L}_t = \mathbf{S}_t - \mathbf{W}_t$ is the Laplacian matrix of $\mathbf{W}_t$. In $\mathbf{W}_t$, each element $w_t(i,j)$ is the indicator of relationship of samples in task t. Specifically, in our task, we follow previous work in [89] to set $w_t(i,j) = 1$ if $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t,j}$ are of the same person across views and $w_t(i,j) = 0$ otherwise. For the target task we initialize $w_t(i,j)$ as all zeros because the target dataset does not provide any label information. The Graph Laplacian term preserves the pairwise relationships of images across camera views. Minimization of $l(\mathbf{A})$ will force the images of the same person across views being closed to each other and therefore enhance the performance of Re-ID .

Domain-aware representation learning. Furthermore, in order to learn the domain-unique appearances in the schema, we aim to emphasize the dissimilarity of representations across different datasets. We introduce a support discriminative term:

$$f(\mathbf{A}) = \sum_{t=1}^{T} \sum_{p} \sum_{q} \| \mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q} \|_0. \tag{6.6}$$

where $\mathbf{a}_{t,p}$ and $\mathbf{a}_{/t,q}$ are the p-th column vector of $\mathbf{A}_t$ and q-th column vector of $\mathbf{A}_{/t}$ receptively. $\mathbf{A}_i$ is the coefficient matrix for the t-th task and $\mathbf{A}_{/t}$ is the sub-matrix of $\mathbf{A}$ with columns of $\mathbf{A}_i$ removed. Here, we let $\mathbf{a} \circ \mathbf{b}$ represent the Hadamard (element-wise) product between two vectors $\mathbf{a}$ and $\mathbf{b}$. Let $\mathbf{a}^{\odot 2} = \mathbf{a} \circ \mathbf{a}$ be the element-wise square of $\mathbf{a}$. The ℓ_0 norm $\| \mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q} \|_0$ presents the number of shared supports of sparse representation $\mathbf{a}_{t,p}$ and $\mathbf{a}_{/t,q}$ of feature vectors $\mathbf{x}_{t,p}$ and $\mathbf{x}_{/t,q}$ [74]. In our task, they represents the camera shots of persons between task t and other tasks. Therefore, minimizing $f(\mathbf{A})$ will decrease the overlapping supports between different datasets, such that will enlarge the dissimilarity of representations between different tasks (datasets). As a result, our algorithm can learn the domain-aware representations simultaneously in dictionary learning.

Finally, we propose our Re-ID algorithm as a unsupervised multi-task dictionary learning optimization problem as following:

$$\begin{aligned}\min_{\mathbf{D},\mathbf{A}} \sum_{t=1}^{T} \{ \|\mathbf{X}_t - \mathbf{D}\mathbf{A}_t\|_F^2 + \alpha \|\mathbf{A}_t\|_{2,1} \\ + \beta Tr(\mathbf{A}_t \mathbf{L}_t \mathbf{A}_t^{\mathsf{T}}) + \gamma \sum_p \sum_q \|\mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q}\|_0 \}.\end{aligned} \tag{6.7}$$

Algorithm 6: Multi-task Dictionary Learning with support discrimination term for Person Re-ID

```
Input: $T$ training Data sets $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_T\}$, regularization leverage parameters $\alpha$, $\beta$ and $\gamma$.
Initialize dictionary $\mathbf{D} \in \mathcal{R}^{d \times k}$, iteration index i = 0.
while not converge do
        Update $\mathbf{A}^{i+1}$ with $\mathbf{D}^i$ according to Algorithm 7.
        Update $\mathbf{D}^{i+1}$ with $\mathbf{A}^{i+1}$ by solving (6.17).
        Update i = i+1;
end while
Output: Dictionary $\mathbf{D}$.
```

Algorithm 7: Sparse Code Learning using ADMM for (6.14)

```
Input: $T$ training Data sets $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_T\}$, regularization leverage parameters $\alpha$, $\beta$ and $\gamma$, penalty parameter $u$, learning step size $\eta$,
Initialize dictionary $\mathbf{D} \in \mathcal{R}^{d \times k}$ with $\mathbf{A}^0 = \mathbf{0}$ and $\Lambda^0 = 0$
for t = 1:T do
    Initialize iteration index i = 0;
    while not converge do
        Compute $\Phi_t^i$ according to (6.10).
        Update $\mathbf{Z}^{i+1} = Shrink(\mathbf{A}_t^i, \Lambda^i, u, \gamma)$.
        Update $\mathbf{A}^{i+1}$ by each of its column $a_t^{i+1}$ according to (6.16).
        Update $\Lambda^{i+1} = \Lambda^i - \eta u(\mathbf{Z}_t^{i+1} - \mathbf{A}_t^{i+1})$.
        Update i = i+1;
    end while
end for
Output: Estimated sparse code $\mathbf{A}$.
```

6.2.2 Optimization

We propose an iterative algorithm to optimize the objective function in (6.7). We describe the optimization algorithm in Algorithm 6. The details of the proposed algorithm are as following:

Optimize A. When fixed $\mathbf{D}$ and $\mathbf{L}_t$, we optimize A_t by solving each task as a subproblem of (6.7):

$$\begin{aligned}\min_{\mathbf{A}_t} \|\mathbf{X}_t - \mathbf{D}\mathbf{A}_t\|_F^2 + \alpha\|\mathbf{A}_t\|_{2,1} \\ + \beta Tr(\mathbf{A}_t\mathbf{L}_t\mathbf{A}_t^{\mathsf{T}}) + \gamma\sum_p\sum_q \|\mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q}\|_0.\end{aligned} \tag{6.8}$$

However, minimizing the last term $\|\mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q}\|_0$ is an NP-hard problem. Following the iterative reweighting schemes in [8, 12, 74, 118], we use the iterative reweighted ℓ_2 minimization to approximate the ℓ_0 norm, which is able to produce more focal estimates in optimization progresses [74]. Specifically, in each iteration $i(i > 1)$ the objective value of $f(\mathbf{A}_t)^{(i)}$ is updated by the reweighted $\ell_{2,1}$-norm $f(\mathbf{A}_t)^{(i)} = \sum\limits_p\sum\limits_q \|\phi_{t,p,q}^{(i)}\mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q}\|_2$, where $\phi_{t,p,q}^{(i)}$ is the weight calculated according to the previous iteration. When updating $f(\mathbf{A}_t)^{(i)}$ in each task t, the vector of other tasks $\mathbf{A}_{/t} = \{\mathbf{a}_{/t,q}\}_{q=1}^{n_t}$ are fixed. In our algorithm, the pairwise weight is estimated as following:

$$\phi_{t,p,q} = \frac{1}{(\mathbf{a}_{t,p}^{'} \circ \mathbf{a}_{/t,q}^{'})^{\odot 2} + \epsilon}. \tag{6.9}$$

where $\mathbf{a}_{t,p}^{'}$ and $\mathbf{a}_{/t,q}^{'}$ are coefficients from the previous iteration. ϵ is a regularization factor decreasing to zero when iteration number increases. Notice computing each pairwise weight is time consuming. Indeed, affected by $\ell_{2,1}$ structure sparsity norm $g(\mathbf{A})$, coefficients in the same task will high probably present a similar sparse structure. Therefore, we approximate the self-squared Hadamard product of each atom by the average of all atoms in the task t as $\forall p$, $(\mathbf{a}_{t,p}^{'})^{\odot 2} \approx (\tilde{\mathbf{a}_t^{'}})^{\odot 2} = \sum_p(\mathbf{a}_{t,p}^{'})^{\odot 2}/n_t$. Thus the approximated weight shared by all atoms in task t is rewritten as $\tilde{\phi}_{t,q} = \frac{1}{(\tilde{\mathbf{a}_t^{'}})^{\odot 2} \circ (\mathbf{a}_{/t,q}^{'})^{\odot 2} + \epsilon}$. We can verify that $\sum\tilde{\phi}_{t,p,q} \circ (\mathbf{a}_{t,p} \circ \mathbf{a}_{/t,q})^{\odot 2} = \operatorname{diag}((\tilde{\phi}_{t,p,q} \circ \mathbf{a}_{/t,q})^{\odot 2} \cdot \mathbf{a}_{t,p}^{\odot 2})$. Overall, we define

$$\Phi_t = \operatorname{diag}(\sqrt{\sum_q(\sqrt{(\tilde{\phi}_{t,q})} \circ \mathbf{a}_{/t,q})^{\odot 2}}), \tag{6.10}$$

and thus $f(\mathbf{A})$ can be rewritten as

$$f(\mathbf{A}) = \sum_{t=1}^{T}\sum_{p}\|\Phi_t\mathbf{a}_{t,p}\|_{2,1} = \sum_{t=1}^{T}\|\Phi_t\mathbf{A}_t\|_F^2. \tag{6.11}$$

Finally (6.8) can be rewritten to

$$\min_{\mathbf{A}_t}\|\mathbf{X}_t-\mathbf{D}\mathbf{A}_t\|_F^2+\alpha\|\mathbf{A}_t\|_{2,1}+\beta Tr(\mathbf{A}_t\mathbf{L}_t\mathbf{A}_t)+\gamma\|\Phi_t\mathbf{A}_t\|_F^2. \tag{6.12}$$

Alternating direction method of multipliers (ADMM) can be used to solve (6.12). First, by introducing an auxiliary variable $\mathbf{Z}_t = \mathbf{A}_t \in \mathbb{R}^{k\times n_t}$, the problem can be reformulated to

$$\begin{aligned}\min_{\mathbf{A}_t,\mathbf{Z}_t}&\|\mathbf{X}_t-\mathbf{D}\mathbf{A}_t\|_F^2+\alpha\|\mathbf{Z}_t\|_{2,1}+\beta Tr(\mathbf{A}_t\mathbf{L}_t\mathbf{A}_t)+\gamma\|\Phi_t\mathbf{A}_t\|_F^2,\\ &s.t.\ \ \mathbf{A}_t-\mathbf{Z}_t=0.\end{aligned} \tag{6.13}$$

Then the augmented Lagrangian function w.r.t $\mathbf{A}_t$ and $\mathbf{Z}_t$ can be formed as

$$\begin{aligned}L_u(\mathbf{A}_t,\mathbf{Z}_t)=&\|\mathbf{X}_t-\mathbf{D}\mathbf{A}_t\|_F^2+\alpha\|\mathbf{Z}_t\|_{2,1}+\beta Tr(\mathbf{A}_t\mathbf{L}_t\mathbf{A}_t)\\ &+\gamma\|\Phi_t\mathbf{A}_t\|_F^2-\mathbf{\Lambda}_t^T(\mathbf{Z}_t-\mathbf{A}_t)+\frac{u_t}{2}\|\mathbf{Z}_t-\mathbf{A}_t\|_2^2,\end{aligned} \tag{6.14}$$

where $\mathbf{\Lambda}_t \in \mathbb{R}^{k\times m}$ is the Lagrangian multipliers and $u_t > 0$ is a penalty parameter. The objective function (6.14) can be minimized by alternately updating $\mathbf{A}_t$ and $\mathbf{Z}_t$. We also use a row shrink function[89] when updating $\mathbf{Z}_t$ which is represented as

$$\mathbf{z}^r = max\{\|\mathbf{q}^r\|_2-\frac{w_t}{u_t},0\}\frac{\mathbf{q}^r}{\|\mathbf{q}^r\|_2},\ r=1,\cdots,k, \tag{6.15}$$

where $\mathbf{q}^r = \mathbf{a}^r + \frac{\lambda_t^r}{u_t}$ and $\mathbf{z}^r$,$\mathbf{a}^r$,$\lambda_{\mathbf{t}}^{\mathbf{r}}$ represent the r-th row of matrix $\mathbf{Z}_t$, $\mathbf{A}_t$, $\mathbf{\Lambda}_t$ respectively. When $\mathbf{Z}_t$ is fixed, we can update $\mathbf{A}_i$ by each column $\mathbf{a}_i$. Optimal solution $\mathbf{a}_t^\star$ can be obtained by setting derivative of L_u w.r.t $\mathbf{a}_t$ to zero, which is similar as in [31, 74, 89]:

$$\begin{aligned}\mathbf{a}_{t,k}^\star=&(\mathbf{D}^T\mathbf{D}+\beta\Phi_t^T\Phi_t+2\beta l_{ii}\mathbf{I}+u_1\mathbf{I})^{-1}\\ &\times(\mathbf{D}^T\mathbf{x}_{t,k}-2\beta\sum_{k\neq i}\mathbf{a}_{t,k}l_{ki}+u_1\mathbf{Z}_i^{k+1}-\frac{1}{2}\lambda_{t,k})\end{aligned} \tag{6.16}$$

where l_{ii} is the (i,i) element of $\mathbf{L}$, $\mathbf{a}_{t,k}$ is the k-th column vector of $\mathbf{A}_t$, $\mathbf{x}_{t,k}$ is the k-th column vector of $\mathbf{X}_t$, $\lambda_{t,k}$ is the k-th column vector of $\mathbf{\Lambda}_t$. The Algorithm solve (6.8) is detailed in Algorithm 7.

Optimize $\mathbf{D}_t$. When given fixed $\mathbf{A}$ and $\mathbf{L}_t$, the optimization problem in (6.7) is equal to

$$\|\mathbf{X}_t - \mathbf{D}\mathbf{A}\|_F^2, \quad s.t. \quad \|\mathbf{d}_i\|_2^2 = 1, \tag{6.17}$$

which is a standard dictionary learning task. We can solve it by updating $\mathbf{d}_i$ column by column. When updating $\mathbf{d}_i$, all the other columns $\mathbf{d}_j$, $j \neq i$ are fixed. Generally, it is required that each column $\mathbf{d}_i$ of $\mathbf{D}$ is a unit vector. It is a quadratic programming problem and it can be solved by using the K-SVD algorithm[97].

Update $\mathbf{W}_t$. Following [89], we update the affinity matrix $\mathbf{W}_t$ the algorithm obtains optimal $\mathbf{A}$ and $\mathbf{D}$. Recall that in the initialization stage, $\mathbf{W}_t$ of source datasets are constructed with labeled information, and $\mathbf{W}_T$ of target dataset in set as zero matrix. After operating Algorithm 7, $\mathbf{W}_t$ for all the source and the target datasets are recomputed with $\mathbf{A}_t$ buy cosine similarity metric $\mathrm{sim}_{\cos}(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{\|\mathbf{a}_i\| \cdot \|\mathbf{a}_j\|}$ where $\mathbf{a}_i$ and $\mathbf{a}_j$ are atoms of coefficient matrix $\mathbf{A}_t$ corresponding to dataset $\mathbf{X}_t$. Specifically, for each $\mathbf{a}_i$, if $\mathbf{a}_j$ is his k-nearest neighborer, the $w_{i,j} = \mathrm{sim}_{\cos}(\mathbf{a}_i, \mathbf{a}_j)$ otherwise $w_{i,j} = 0$. In our work we set $k = 5$. With the renewed $\mathbf{W}_t$, we re-run Algorithm 7 in the next criterion. The termination condition is set as an loose stopping criterion when $\frac{|Lu^{k+1} - Lu^k|}{|Lu^0|} \leq \varepsilon$. In practice, ε is set to 0.1 and the total iteration number is typically under 5 in our experiments.

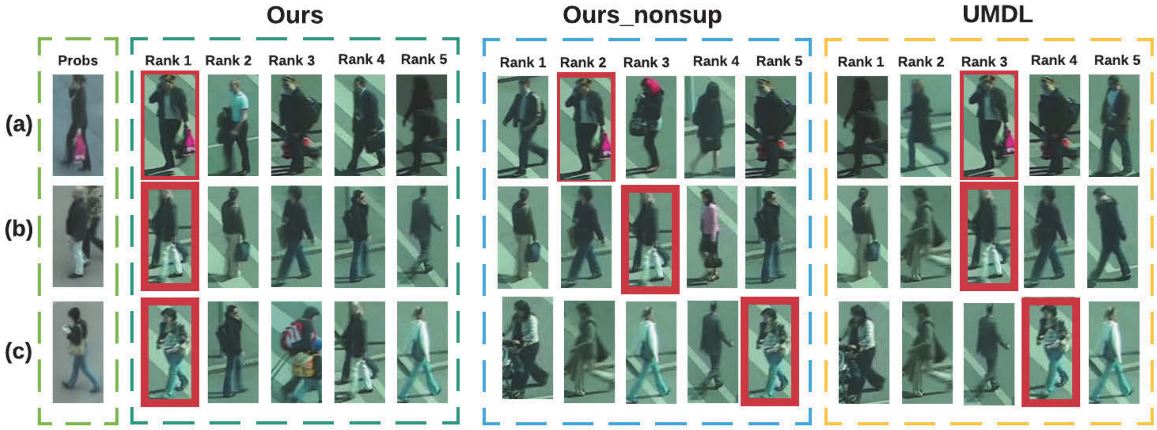


Figure 6.3 Performances of UDML and Ours algorithm on PRID dataset. Probe images are provided in the left column. Top-5 candidate images are sorted in descent order according to their score. The ground truth images are marked with red bounding boxes. (a) Woman with pink handbag and dark coat. (b) Man with dark coat and white pants (and without handbag). (c) Woman with white handbag and dark coat.

6.3 Experiments

6.3.1 Experimental Settings

Datasets. We compare our algorithm with the state-of-the-art algorithms on four widely referred benchmark dataset of Peron Re-identification. The **VIPeR** [30] dataset contains 1,264 images of 632 persons from two non-overlapping camera views. Two images are taken for each person, each from a different camera. Viewpoint changes and varying illumination conditions have occurred. The **PRID** [35] dataset contains images of 385 individuals from two distinct cameras. Camera B records 749 persons and Camera A records 385 persons, with 200 of them are same persons. The **iLIDS** [139] dataset records 119 individuals captures by three different cameras in an airport terminal. It contains 476 images with large occlusions caused by luggage and viewpoint changes. The **CAVIAR** [14] dataset for Re-ID contains 72 individuals captured by two cameras in a shopping mall. The amount of image is 1,220, with 10 to 20 images for each individual. The size of images in the CAVIAR dataset vary significantly from 39×17 to 141×72. The **CUHK03** [61] dataset contains images of 1360 different individuals captured in a campus by six cameras. A total number of 13,164 images are recorded and 4.8 images are recorded for each individual on everage.

As in [69, 89], we scaled all images to 128×48 pixel images and normalized to color+HOG+LBP histogram-based 5138-D feature representations [67]. The size of dictionaries is set to 150 for all experiments. Other parameters are tuned by four-fold cross-validation method.

Algorithms. We first consider the single-task experiments as a baseline. In the single-task experiments, there is no source data for transfer learning. Therefore, we could invesgate wether the cross-data transfer learning (the multi-task methods) can improve the performance of unsupervised Re-ID . (a) *Single-task methods:* **SDALF:** [21] SDALF generates hand-crafted-feature for unsupervised Re-ID learning by exploiting the property of symmetry in pedestrian images. **eSDC:** [135] eSDC method introduces unsupervised saliency learning for Re-ID task. **GTS:** [113] GTS is proposed to solve Re-ID problem by exploring generative probabilistic topic modeling. **ISR:** [68] ISR applies sparse representation recognition model for Re-ID built on sparse basis expansions. **CAMEL:** [129] CAMEL introduces a cross-view asymmetric metric learning for unsupervised Re-ID . **UMDL_S:** We also involve unsupervised multi-task dictionary method UMDL [89] method with no source data related term for single task test, which is denoted as UMDL_S. **Ours_S:** For single-task experiment, there is only one dataset. Therefore, the support discriminative term and the Graph Laplacian term are not activated and only the structure sparse term is used.

(b) *Multi-task methods:* There are few unsupervised cross-dataset multi-task learning applied for person re-identification. As in former works [89], we additionally invite several multi-task learning methods as baselines : **AdaRSVM:**[78] AdaRSVM is a cross-domain unsupervised adaptive ranking SVM learning method designed for person re-identification. It permit the information of negative pairs in target training. **SA_DA+kLFDA:**(abbr. SA+kLFDA) In the framework, SA_DA is an unsupervised domain adaptation algorithm that aligns the source and target domain through data distributions. After domain adaptation, the supervised Re-ID algorithm kLFDA is implemented on labeled source data and then applied to the aligned target dataset [89]. **kLFDA_N:** Furthermore, we provide a transfer learning method baseline. In the framework, kLFDA algorithm is first trained on source datasets and applied directly to target dataset with no model adaptation. We denoted this algorithm as kLFDA_N. **UMDL:** In [89], the authors proposed a multi-task cross-dataset dictionary learning algorithm with a Laplacian regularization term for person re-identification. **Ours:** In the experiments, we denote our algorithm as Ours. For ablation study, we also perform our algorithms without support discriminative term in the test, which is denoted as **Ours_nonsup**.

Settings. In each experiment, one dataset is selected as the target dataset, the other datasets are chosen as the source datasets. For the target dataset, no label information is utilized in training stage; For the source datasets, label information is utilized for initializing the corresponding Laplacian matrix as mentioned in (6.5).We report the average performance of 20 independent trials. In each trial, we randomly divide each dataset into two equal-sized subsets as training and testing sets, with no overlapping on person identities. For datasets recording two camera views, e.g. VIPeR, PRID, images from one view are selected randomly as probe sets, and images from other views are chosen as gallery sets. For multi-view dataset, e.g. iLIDS, one view is selected randomly as gallery images with others are chosen as probe sets.

6.3.2 Experimental Analysis

Performance of Person Re-identification

As Shown in Table. 6.1, we display the performance of the investigated algorithms by rank one machining accuracy(%). We observe that: (1) In multi-task learning, our algorithm consistently shown outstanding performance on all the datasets. Specifically,

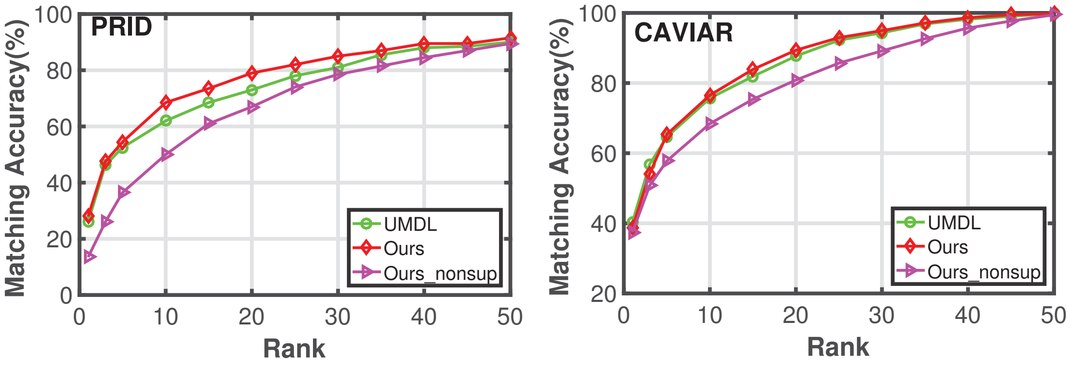
Table 6.1 Rank One Matching Accuracy(%) on Unsupervised Re-ID. (a) Single-task methods. (b) Multi-task methods.

	Dataset	VIPeR	PRID	CAVIAR	iLIDS	CUHK03
(a)	SDALF	19.9	16.3	-	29.0	-
	eSDC	26.7	-	-	36.8	8.76
	GTS	25.2	-	-	42.4	-
	ISR	27.0	17.0	29.0	39.5	11.46
	CAMEL	30.9	-	-	-	31.9
	UMDL_S	24.3	14.1	33.5	45.7	13.8
	OURS_S	26.9	19.2	31.6	44.3	16.2
(b)	kLFDA_N	12.9	8.5	32.8	36.9	7.6
	SA+kLFDA	11.6	8.1	32.1	35.8	6.8
	AdaRSVM	10.9	4.9	5.8	-	5.8
	UMDL	31.5	24.2	41.6	49.3	27.1
	OURS_nonsup	24.7	22.52	40.1	48.1	26.8
	OURS	**31.9***	**27.9***	**42.5***	**50.3***	**35.2***

our algorithm and UMDL algorithm performance much better than other algorithms. It indicates that the cross-data dictionary learning can improve the performance of Re-ID . Moreover, our algorithm presents better performance on all the datasets than UMDL. Especially, our algorithm outperforms others by up to 5% in the experiments. (2) Compared to single-task learning, UMDL and our algorithm report a much better performance than UMDL_S and Ours_S. It indicates that utilizing the knowledge from source domains can improve the Re-ID performance. (3) In multi-task learning, the results of kLFDA_N reports weaker performance, which indicates directly transfer knowledge from source datasets to target dataset do not help much in Re-ID . Meanwhile, SA_DA+kLFDA does not report high performance, too. It indicates that unsupervised domain adaptation methods may not be suitable to cross-datasets Re-ID . The reason is domain adaptation methods assume that all domains have the same classification tasks but in our Re-ID problem classes of persons are completely different as the persons are non-overlapped.

Ablation Study

For the ablation sutudy, we analyze performances of our algorithms (Ours) and our algorithm without the support discriminative term (Ours_nonsup). UMDL is also invited as a baseline as it is also a cross-dataset algorithm repor the state-of-the-art performance in previous works.



(a) PRID

(b) CAVIAR

Figure 6.4 Re-id performance for ablation study.

(1) Numerical analysis First, We run Ours, Ours_nonsup and UMDL on on PRID and CAVIAR datasets. As shown in Fig.6.4(a) and Fig.6.4(b), the Ours algorithm presents the best performance over UMDL and Ours_nonsup on matching accuracy consistently on the two datasets. It outperforms others very early around 5% from top-10 rank performance on PRID and 3% on CAVIAR until all algorithms meet over 90% matching accuracy on top-50 rank performance. It implies that introducing support discriminative term improves the Re-ID performance.

(2) Case analysis As we shown earlier in Fig. 6.1, common appearances and domain-unique appearance are observed in Re-ID datasets. In this section, we aim to test the ability of our algorithms in discovering these domain-unique appearances. We pick up three persons with domain-unique appearances as probe images and select top-5 ranked candidates from the gallery sets the tested algorithms. Candidates are ranked according to their similarities to the probe in descending order. Ground truth person images are marked with red bounding boxes.

As displayed in Fig. 6.3, UMDL prefers to rank candidates based on the common appearances over the domain-unique appearances. For the person (a), UMDL weights person with 'dark coat and pants' over 'carrying pink handbag'; For the person (b), it orders the person with 'dark color coat and white pants' over 'carrying handbags (or not)', such that the imposters wearing the similar clothes but carrying handbags are ranked over the ground truth person, which does not carrying bags at all. For the person (c) who is taking white bags, because of a heavily viewpoint change, UMDL mismatches the probe to imposters with white coats. Meanwhile,our algorithm can learn the domain-unique appearances and thus successfully matches the probes to the correct candidates. In the top-5 ranked candidates, we observe that our algorithm can select candidates with domain-unique appearances. For the person (a), it selects four candidates carrying handbags; for the person (c), there are four candidates carrying shoulder bags.

Moreover, we could verify that the support discriminative term emphasizes the diversity of unique appearance. As shown in Fig. 6.1, if there is no support discriminative term, our algorithm (denoted as Ours_nonsup) will not be able to recognize the domain-unique feature. The performance of Ours_nonsup degrades dramatically as shown in Table 6.1.

(3) Inter-dataset similarity analysis In order to verify the ability of our algorithm on discriminative representation learning, we further calculate the *similarity index* of samples inter-datasets. Idealy, higher similarity index indicates that the features of samples between two datasets are more similar. In particular, we define the similarity index of dataset $\mathbf{A}_i$ to dataset $\mathbf{A}_j$ as: $S_{ij} = \sum_p \sum_q \|\mathbf{a}_{i,p} \odot \mathbf{a}_{j,q}\|_0 / n_i$ where $\mathbf{a}_{i,p}$ is the p-th column of $\mathbf{A}_i$ and $\mathbf{a}_{j,q}$ is the q-th column of $\mathbf{A}_j$. Experiments are re-run on PRID dataset with the best performed parameter $\alpha = 1$, $\beta = 10^{-3}$,$\gamma = 3$. We display the inter-dataset similarity index in Table. 6.2. As shown in the table, UMDL and Ours_nonsup algorithm performances distinctly higher similarity indexes cross datasets w.r.t. PRID then Ours algorithm. It implies the support discriminative term can enhance the discriminative capability of representations across-datasets.

Table 6.2 Inter-dataset Similarity Index with Target Dataset PRID ($\times 10^2$)

	UMDL	Ours_nonsup	Ours
VIPeR w.r.t. PRID	12.50	8.24	0.29
CAVIAR w.r.t. PRID	13.01	7.59	0.27
iLID w.r.t. PRID	12.94	10.59	0.22
CUHK03 w.r.t. PRID	19.33	15.49	0.43

6.4 Summary

In this chapter, we proposed a novel domain-aware unsupervised cross-dataset approach on person re-identification. It is able to characterize both the shared and domain-unique representations cross different camera-view network domains. Experiments on four real-world datasets consistently report outstanding performance of our algorithm. Analysis on selected cases also support that our algorithm enhances the Re-ID performance by utilizing domain-unique representations. Future works are suggested to extend the proposed algorithm to real-world Re-ID scenario with more source domains with various domain-unique appearances.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This work addresses the problem of visual analysis with limit supervision. In the work, the limitation supervision is categorized in three scenarios, based on the different stages of experiments, the accessibility of labeled data and the limitation of annotation resource. Multiple algorithms are proposed for each scenario:

Chapter 2 and Chapter 3 attempt to solve the problem when there is no labeled data at the early stage of the experiments and annotation resource is limited. First, Early Active Learning with Pairwise Constraint is proposed. It is the first instance-based early active learning method that is applied to visual retrieval task for person Re-ID . The pairwise constraint is introduced to capture the relativeness of instances of data for annotation. Second, Pair-based Early Active Learning is proposed to select the most informative pairs of samples to annotation. Meanwhile, the diversity of the pairs are considered in the schema to enhance representativeness of labeled data. The experimental results confirm the effectiveness of the proposed approach in comparison to multiple active learning algorithms on several datasets.

In Chapter 4, we address the visual retrieval task with scarce labeled data and abundant unlabeled data. A semi-supervised attribute learning algorithm is proposed. It jointly learns the latent attributes with appropriate dimensions and estimates the pairwise probability of the data simultaneously. The experimental results confirm the effectiveness of the proposed approach.

In Chapter 5 and Chapter 6, transfer learning is studied to complete the visual analysis task. Our methods utilize and learn prior knowledge from source domains with sufficient labeled data and transfers such knowledge to the target domain. First, to address the unsupervised visual retrieval task in person Re-ID , an Analogical Transfer

Learning schema is proposed for this problem. It attempts to select only the helpful source domain instances to enhance the models of the target task. Second, the few-shot visual classification problem is addressed through a Domain-aware Unsupervised Cross-dataset Transfer Learning algorithm. In the algorithm, the importance of common and domain-unique appearances are estimated simultaneously and jointly contribute to the representation learning in the visual classification task. The experimental results show that our algorithms are effective on both person Re-ID and image classification tasks.

7.2 Future Work

In this work, we clarify that difficulties of visual analysis vary in the different stage of experiments. For the future work, the improved approaches can be considered in the same directions:

First, our early active learning algorithms can be applied to other applications such as recommender systems [96]. Furthermore, our algorithms can improve the deep active learning methods, such as in face recognition [65] and image classification [116].

Second, for the semi-supervised attribute learning, our algorithm only considers the sample pairwise probabilities. In the future, it is highly suggested to be extended to triplet and quadruplet relationship analysis [13], which is widely discussed in person Re-ID . Moreover, deep relation learning framework [115] can be introduced to our schema for further improvements.

Third, the proposed analogical transfer learning can be adopted to solve the deep feature representation learning problem with deep neural network frameworks [120]. For the Domain-aware Unsupervised Cross-dataset Transfer Learning algorithm, future works can extend and improve the algorithm in unsupervised domain adaption with deep learning frameworks [76].

Bibliography

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

[2] Ao, S., Li, X., and Ling, C. X. (2017). Effective multiclass transfer for hypothesis transfer learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 64–75. Springer.

[3] Aytar, Y. and Zisserman, A. (2011). Tabula rasa: Model transfer for object category detection. In *2011 International Conference on Computer Vision*, pages 2252–2259. IEEE.

[4] Balcan, M.-F., Broder, A., and Zhang, T. (2007). Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer.

[5] Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, pages 81–88. ACM.

[6] Broderick, T., Kulis, B., and Jordan, M. (2013). Mad-bayes: Map-based asymptotic derivations from bayes. In *International Conference on Machine Learning*, pages 226–234.

[7] Bruzzone, L. and Marconcini, M. (2010). Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787.

[8] Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted l 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905.

[9] Caruana, R. (1998). Multitask learning. In *Learning to Learn*, pages 95–133. Springer.

[10] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[11] Chang, X., Yu, Y., Yang, Y., and Xing, E. P. (2017). Semantic pooling for complex event analysis in untrimmed videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1617–1632.

[12] Chartrand, R. and Yin, W. (2008). Iteratively reweighted algorithms for compressive sensing. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pages 3869–3872. IEEE.

[13] Chen, W., Chen, X., Zhang, J., and Huang, K. (2017). Beyond triplet loss: a deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

[14] Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*.

[15] Choi, J., Hwang, S. J., Sigal, L., and Davis, L. S. (2016). Knowledge transfer with interactive learning of semantic relationships. In *AAAI Conference on Artificial Intelligence*.

[16] Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the International Conference on Machine Learning*, pages 193–200. ACM.

[17] Delbos, F. and Gilbert, J. C. (2003). *Global linear convergence of an augmented Lagrangian algorithm for solving convex quadratic optimization problems*. PhD thesis, INRIA.

[18] Deng, Y., Luo, P., Loy, C. C., and Tang, X. (2014). Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM.

[19] Ding, Z. and Fu, Y. (2016). Robust transfer metric learning for image classification. *IEEE Transactions on Image Processing*, 26(2):660–670.

[20] Fang, M., Yin, J., Zhu, X., and Zhang, C. (2015). Trgraph: Cross-network transfer learning via common signature subgraphs. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2536–2549.

[21] 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. IEEE.

[22] Feng, J., Jegelka, S., Yan, S., and Darrell, T. (2014). Learning scalable discriminative dictionary with sample relatedness. In *CVPR*, pages 1645–1652.

[23] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168.

[24] Gao, J., Fan, W., Jiang, J., and Han, J. (2008). Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 283–291. ACM.

[25] Gärdenfors, P. (2003). *Belief revision*, volume 29. Cambridge University Press.

[26] Ge, L., Gao, J., Ngo, H., Li, K., and Zhang, A. (2014). On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Statistical Analysis and Data Mining*, 7(4):254–271.

[27] Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

[28] Gönen, M. and Margolin, A. A. (2014). Kernelized bayesian transfer learning. In *AAAI Conference on Artificial Intelligence*, pages 1831–1839.

[29] Gong, C., Tao, D., Maybank, S. J., Liu, W., Kang, G., and Yang, J. (2016). Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260.

[30] Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3.

[31] Guo, H., Jiang, Z., and Davis, L. S. (2012). Discriminative dictionary learning with pairwise constraints. In *Asian Conference on Computer Vision*, pages 328–342. Springer.

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

[33] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

[34] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM.

[35] Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer.

[36] Hirzer, M., Roth, P. M., Köstinger, M., and Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. Springer.

[37] Hoffman, J., Rodner, E., Donahue, J., Darrell, T., and Saenko, K. (2013). Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*.

[38] Hou, C.-A., Tsai, Y.-H. H., Yeh, Y.-R., and Wang, Y.-C. F. (2016). Unsupervised domain adaptation with label and structural consistency. *IEEE Transactions on Image Processing*, 25(12):5552–5562.

[39] Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

[40] Huang, S.-J., Jin, R., and Zhou, Z.-H. (2010). Active learning by querying informative and representative examples. In *Advances in Neural Information Processing Systems (NIPS)*, pages 892–900.

[41] Jaakkola, T., Meila, M., and Jebara, T. (2000). Maximum entropy discrimination. In *Advances in neural information processing systems*, pages 470–476.

[42] Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the annual meeting on Association for Computational Linguistics*, volume 7, pages 264–271.

[43] Jiang, L., Meng, D., Yu, S.-I., Lan, Z., Shan, S., and Hauptmann, A. (2014). Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, pages 2078–2086.

[44] Jiang, L., Meng, D., Zhao, Q., Shan, S., and Hauptmann, A. G. (2015). Self-paced curriculum learning. In *AAAI Conference on Artificial Intelligence*, volume 2, page 6.

[45] Joachims, T. (2006). Transductive support vector machines. *Chapelle et al.(2006)*, pages 105–118.

[46] Kale, D., Ghazvininejad, M., Ramakrishna, A., He, J., and Liu, Y. (2015). Hierarchical active transfer learning. In *Society for Industrial and Applied Mathematics Publications*. SIAM.

[47] Karanam, S., Li, Y., and Radke, R. J. (2015a). Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *CVPR*, pages 4516–4524.

[48] Karanam, S., Li, Y., and Radke, R. J. (2015b). Sparse re-id: Block sparsity for person re-identification. In *CVPR*, pages 33–40.

[49] Kodirov, E., Xiang, T., Fu, Z., and Gong, S. (2016). Person re-identification by unsupervised l 1 graph learning. In *European Conference on Computer Vision*, pages 178–195. Springer.

[50] Kodirov, E., Xiang, T., and Gong, S. (2015). Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *British Machine Vision Conference (BMVC)*, volume 3, page 8.

[51] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE.

[52] Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197.

[53] Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *Proceedings of the International Conference on Machine Learning*, pages 942–950.

[54] Kuzborskij, I. and Orabona, F. (2014). Fast rates by transferring from auxiliary hypotheses. *arXiv preprint arXiv:1412.1619*.

[55] Kuzborskij, I., Orabona, F., and Caputo, B. (2013). From n to n+ 1: Multiclass transfer incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365.

[56] Kuzborskij, I., Orabona, F., and Caputo, B. (2014). Scalable greedy algorithms for transfer learning. *arXiv preprint arXiv:1408.1292*.

[57] Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the International Conference on Machine Learning*, pages 331–339.

[58] Lewis, D. D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156.

[59] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

[60] Li, S., Shao, M., and Fu, Y. (2015). Cross-view projective dictionary learning for person re-identification. In *IJCAI*, pages 2155–2161.

[61] 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.

[62] 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.

[63] Liao, S. and Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. In *CVPR*, pages 3685–3693.

[64] Liao, X., Xue, Y., and Carin, L. (2005). Logistic regression with an auxiliary data source. In *Proceedings of the International Conference on Machine Learning*, pages 505–512. ACM.

[65] Lin, L., Wang, K., Meng, D., Zuo, W., and Zhang, L. (2018). Active self-paced learning for cost-effective and progressive face identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):7–19.

[66] 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*.

[67] Lisanti, G., Masi, I., Bagdanov, A., and Del Bimbo, A. (2014a). Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1.

[68] Lisanti, G., Masi, I., Bagdanov, A. D., and Del Bimbo, A. (2015). Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642.

[69] Lisanti, G., Masi, I., and Del Bimbo, A. (2014b). Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*, page 10. ACM.

[70] Liu, F., Xu, X., Qiu, S., Qing, C., and Tao, D. (2016a). Simple to complex transfer learning for action recognition. *IEEE Transactions on Image Processing*, 25(2):949–960.

[71] Liu, T., Tao, D., Song, M., and Maybank, S. J. (2017a). Algorithm-dependent generalization bounds for multi-task learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):227–241.

[72] Liu, W., Chang, X., Chen, L., and Yang, Y. (2017b). Early active learning with pairwise constraint for person re-identification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 103–118. Springer.

[73] 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*, pages 3550–3557.

[74] Liu, Y., Chen, W., Chen, Q., and Wassell, I. (2016b). Support discrimination dictionary learning for image classification. In *ECCV*, pages 375–390. Springer.

[75] Long, M., Wang, J., Sun, J., and Philip, S. Y. (2015). Domain invariant transfer kernel learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1519–1532.

[76] Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016). Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144.

[77] Lv, J., Chen, W., Li, Q., and Yang, C. (2018). Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. *arXiv preprint arXiv:1803.07293*.

[78] Ma, A. J., Li, J., Yuen, P. C., and Li, P. (2015). Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Transactions on Image Processing*, 24(5):1599–1613.

[79] Ma, A. J. and Li, P. (2014). Semi-supervised ranking for re-identification with few labeled image pairs. In *Asian Conference on Computer Vision*, pages 598–613. Springer.

[80] Martinel, N., Foresti, G. L., and Micheloni, C. (2016). Person reidentification in a distributed camera network framework. *IEEE transactions on cybernetics*.

[81] Nguyen, H. T. and Smeulders, A. (2004). Active learning using pre-clustering. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM.

[82] Nguyen, M. L., Tsang, I. W., Chai, K. M. A., and Chieu, H. L. (2014). Robust domain adaptation for relation extraction via clustering consistency. In *Proceedings of the annual meeting on Association for Computational Linguistics*, pages 807–817.

[83] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

[84] Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). Efficient and robust feature selection via joint l2, 1-norms minimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1813–1821.

[85] Nie, F., Wang, H., Huang, H., and Ding, C. H. (2013). Early active learning via robust representation and structured sparsity. In *International joint conference on artificial intelligence(IJCAI)*.

[86] Nie, F., Xu, D., and Li, X. (2012). Initialization independent clustering with actively self-training method. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):17–27.

[87] Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., and Sandini, G. (2009). Model adaptation with least-squares svm for adaptive hand prosthetics. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 2897–2903. IEEE.

[88] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[89] Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., and Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315.

[90] Pirsiavash, H., Ramanan, D., and Fowlkes, C. C. (2009). Bilinear classifiers for visual recognition. In *Advances in neural information processing systems*, pages 1482–1490.

[91] Prosser, B. J., Zheng, W.-S., Gong, S., Xiang, T., and Mary, Q. (2010). Person re-identification by support vector ranking. In *British Machine Vision Conference (BMVC)*, volume 2, page 6.

[92] Qiu, Q., Ni, J., and Chellappa, R. (2014). Dictionary-based domain adaptation methods for the re-identification of faces. In *Person Re-Identification*, pages 269–285. Springer.

[93] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

[94] Ravi, S. and Larochelle, H. (2016). Optimization as a model for few-shot learning.

[95] Ren, C.-X., Dai, D.-Q., Huang, K.-K., and Lai, Z.-R. (2014). Transfer learning of structured representation for face recognition. *IEEE Transactions on Image Processing*, 23(12):5440–5454.

[96] Rubens, N., Elahi, M., Sugiyama, M., and Kaplan, D. (2015). Active learning in recommender systems. In *Recommender systems handbook*, pages 809–846. Springer.

[97] Rubinstein, R., Zibulevsky, M., and Elad, M. (2008). Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. *Cs Technion*, 40(8):1–15.

[98] Schumann, A. and Stiefelhagen, R. (2017). Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1435–1443. IEEE.

[99] Seah, C.-W., Tsang, I. W.-H., and Ong, Y.-S. (2011). Healing sample selection bias by source classifier selection. In *2011 IEEE 11th International Conference on Data Mining*, pages 577–586. IEEE.

[100] Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 287–294. ACM.

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

[102] Snell, J., Swersky, K., and Zemel, R. S. (2017). Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

[103] Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L. S., and Gao, W. (2015). Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, pages 3739–3747.

[104] 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. Springer.

[105] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440.

[106] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2017). Learning to compare: Relation network for few-shot learning. *arXiv preprint arXiv:1711.06025*.

[107] Supancic, J. S. and Ramanan, D. (2013). Self-paced learning for long-term tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2379–2386.

[108] Tao, D., Jin, L., Wang, Y., and Li, X. (2015). Person reidentification by minimum classification error-based kiss metric learning. *IEEE transactions on Cybernetics*, 45(2):242–252.

[109] Tommasi, T., Orabona, F., and Caputo, B. (2010). Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3081–3088. IEEE.

[110] Tommasi, T., Orabona, F., and Caputo, B. (2014). Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):928–941.

[111] Twomey, N., Diethe, T., and Flach, P. (2015). Bayesian active learning with evidence-based instance selection. In *Workshop on Learning over Multiple Contexts, European Conference on Machine Learning (ECML'15)*.

[112] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638.

[113] Wang, H., Gong, S., and Xiang, T. (2014a). Unsupervised learning of generative topic saliency for person re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press.

[114] Wang, H., Nie, F., and Huang, H. (2014b). Robust distance metric learning via simultaneous l1-norm minimization and maximization. In *International Conference on Machine Learning*, pages 1836–1844.

[115] Wang, H., Shi, X., and Yeung, D.-Y. (2017a). Relational deep learning: A deep latent variable model for link prediction. In *AAAI*, pages 2688–2694.

[116] Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. (2017b). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600.

[117] Wang, Y.-X. and Hebert, M. (2016). Learning by transferring from unsupervised universal sources. In *AAAI Conference on Artificial Intelligence*.

[118] Wipf, D. and Nagarajan, S. (2010). Iterative reweighted l1 and l2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329.

[119] Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2017). Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*.

[120] 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.

[121] Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528.

[122] Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16. Springer.

[123] Xu, C., Tao, D., and Xu, C. (2015). Multi-view self-paced learning for clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press*, pages 3974–3980.

[124] Xu, M., Zhu, J., and Zhang, B. (2013). Fast max-margin matrix factorization with data augmentation. In *International Conference on Machine Learning*, pages 978–986.

[125] Xu, Y., Pan, S. J., Xiong, H., Wu, Q., Luo, R., Min, H., and Song, H. (2017). A unified framework for metric transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(6):1158–1171.

[126] Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1).

[127] Yang, L., Jin, R., and Sukthankar, R. (2012). Bayesian active distance metric learning. *arXiv preprint arXiv:1206.5283*.

[128] Yang, Y., Shen, H. T., Ma, Z., Huang, Z., and Zhou, X. (2011). l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

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

[130] Yu, K., Bi, J., and Tresp, V. (2006). Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM.

[131] Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q., and Han, J. (2015). A self-paced multiple-instance learning framework for co-saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 594–602.

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

[133] Zhang, Y. and Yeung, D.-Y. (2012). A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*.

[134] Zhao, Q., Meng, D., Jiang, L., Xie, Q., Xu, Z., and Hauptmann, A. G. (2015). Self-paced learning for matrix factorization. In *AAAI Conference on Artificial Intelligence*, pages 3196–3202.

[135] Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

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

[137] 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.

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

[139] Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., and Cai, D. (2011). Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336.

[140] Zheng, Z., Zheng, L., and Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *CVPR*.

[141] Zhou, H., Ithapu, V. K., Ravi, S. N., Singh, V., Wahba, G., and Johnson, S. C. (2016). Hypothesis testing in unsupervised domain adaptation with applications in neuroscience. In Lee, D. D., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances In Neural Information Processing Systems*, pages 2496–2504. Curran Associates, Inc.