Deep Multi-View Learning for Visual Understanding

Xiaobin Chang

Submitted in partial fulfilment of the requirement for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary University of London

7 December 2019

To my beloved parents.

Deep Multi-View Learning for Visual Understanding

Xiaobin Chang

Abstract

Multi-view data is the result of an entity being perceived or represented from multiple perspectives. Plenty of applications in visual understanding contain multi-view data. For example, the face images for training a recognition system are usually captured by different devices from multiple angles. This thesis focuses on the cross-view visual recognition problems, e.g., identifying the face images of the same person across different cameras. Several representative multi-view settings, from the supervised multi-view learning to the more challenging unsupervised domain adaptive (UDA) multi-view learning, are investigated. Novel multi-view learning algorithms are proposed correspondingly. To be more specific, the proposed methods are based on the advanced deep neural network (DNN) architectures for better handling visual data. However, directly combining the multi-view learning objectives with DNN can result in different issues, e.g., on scalability, and limit the application scenarios and model performance. Corresponding novelties in DNN methods are thus required to solve them. This thesis is organised into three parts. Each chapter focuses on a multi-view learning setting with novel solutions and is detailed as follows:

Chapter 3 A supervised multi-view learning setting with two different views are studied. To recognise the data samples across views, one strategy is aligning them in a common feature space via correlation maximisation. It is also known as canonical correlation analysis (CCA). Deep CCA has been proposed for better performance with the non-linear projection via deep neural networks. Existing deep CCA models typically decorrelate the deep feature dimensions of each view before their Euclidean distances are minimised in the common space. This feature decorrelation is achieved by enforcing an exact decorrelation constraint which is computationally expensive due to the matrix inversion or SVD operations. Therefore, existing deep CCA models are inefficient and have scalability issues. Furthermore, the exact decorrelation is incompatible with the gradient based deep model training and results in sub-optimal solution. To overcome these aforementioned issues, a novel deep CCA model *Soft CCA* is introduced in this thesis. Specifically, the exact decorrelation is replaced by soft decorrelation via a mini-batch based *Stochastic Decorrelation Loss (SDL)*. It can be jointly optimised with the other training objectives. In addition, our SDL loss can be applied to other deep models beyond multi-view learning.

Chapter 4 The supervised multi-view learning setting, whereby more than two views exist, are studied in this chapter. Recently developed deep multi-view learning algorithms either learn a latent visual representation based on a single semantic level and/or require laborious human annotation of these factors as attributes. A novel deep neural network architecture, called Multi-Level Factorisation Net (MLFN), is proposed to automatically factorise the visual appearance into latent discriminative factors at multiple semantic levels without manual annotation. The main purpose is forcing different views share the same latent factors so that they are can be aligned at all layers. Specifically, MLFN is composed of multiple stacked blocks. Each block

contains multiple factor modules to model latent factors at a specific level, and factor selection modules that dynamically select the factor modules to interpret the content of each input image. The outputs of the factor selection modules also provide a compact latent factor descriptor that is complementary to the conventional deeply learned feature, and they can be fused efficiently. The effectiveness of the proposed MLFN is demonstrated by not only the large-scale cross-view recognition problems but also the general object categorisation tasks.

Chapter 5 The last problem is a special unsupervised domain adaptation setting called unsupervised domain adaptive (UDA) multi-view learning. It contains a fully annotated dataset as the source domain and another unsupervised dataset with relevant tasks as the target domain. The main purpose is to improve the performance of the unlabelled dataset with the annotated data from the other dataset. More importantly, this setting further requires both the source and target domains are multi-view datasets with relevant tasks. Therefore, the assumption of the aligned label space across domains is inappropriate in the UDA multi-view learning. For example, the person re-identification (Re-ID) datasets built on different surveillance scenarios are with images of different people captured and should be given disjoint person identity labels. Existing methods for UDA multi-view learning problems are aligning different domains either in the raw image space or a feature embedding space for domain alignment. In this thesis, a different framework, multi-task learning, is adopted with the domain specific objectives for a common space learning. Specifically, such common space is proposed to enable the knowledge transfer. The conventional supervised losses can be used for the labelled source data while the unsupervised objectives for the target domain play the key roles in domain adaptation. Two novel unsupervised objectives are introduced for UDA multi-view learning and result in two models as below.

The first model, termed *common factorised space model (CFSM)*, is built on the assumptions that the semantic latent attributes are shared between the source and target domains since they are relevant multi-view learning tasks. Different from the existing methods that based on domain alignment, CFSM emphasizes on transferring the information across domains via discovering discriminative latent factors in the proposed common space. However, the multi-view data from target domain is without labels. Therefore, an unsupervised factorisation loss is derived and applied on the common space for latent factors discovery across domains.

The second model still learns a shared embedding space with multi-view data from both domains but with a different assumption. It attempts to discover the latent correspondence of multi-view data in the unsupervised target data. The target data's contribution comes from a clustering process. Each cluster thus reveals the underlying cross-view correspondences across multiple views in target domain. To this end, a novel *Stochastic Inference for Deep Clustering (SIDC)* method is proposed. It reduces self-reinforcing errors that lead to premature convergence to a sub-optimal solution by changing the conventional deterministic cluster assignment to a stochastic one.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged. Some works have been published or are under review as follows:

Chapter 3

1. Xiaobin Chang, Tao Xiang, and Timothy M. Hospedales. *Scalable and effective deep cca via soft decorrelation.* In Proc. of the IEEE Conference on Computer Vision and Pattern, Salt Lake City, Utah, USA, June 2018. (CVPR).

Chapter 4

1. Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. *Multi-level factorisation net for person re-identification.* In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, June 2018. (CVPR).

Chapter 5

1. Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M. Hospedales. *Disjoint Label Space Transfer Learning with Common Factorised Space.* In Proc. of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, January 2019. (AAAI).

2. Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M. Hospedales. *Unsupervised Domain Adaptive Person Re-Identification with Deep Clustering.* Under Review of Proc. of the IEEE International Conference on Computer Vision (ICCV) 2019.

Acknowledgements

First of all, I would like to convey my sincere gratitude to my supervisor, Professor Tao (Tony) Xiang. During my four-year study, Tony gives me plenty of valuable advice and unconditional supports. At the same time, I appreciate the detailed guidance on my research from my co-supervisor Dr. Timothy Hospedales. Finally, many thanks to Professor Shaogang (Sean) Gong, who gave me the advice and supports I needed. With their helps and supervisions, I gradually learn how to conduct research independently.

My friends and colleagues are always willing to help. Dr. Xiatian (Eddy) Zhu has the very comprehensive understanding on different research areas and gave me patient explanations when I have doubts. I also learned plenty of deep insights on research from Dr. Yongxin Yang. The system staff, especially Tim Kay and Haris Krikelis, provide me the strongest technical supports. Melissa Yeo is patient to answer all my questions on administrative processes. I will miss the nice coffee and snacks from June Coster. Special thanks to many other colleagues and friends at Queen Mary. The happy moments we had are one of the most precious treasures in my life.

Last but not least, I am really grateful for the endless and selfless love and supports from my family, especially my parents and grandparents. Moreover, in deep memory of my loving grandmother. It is my turn to carry out more duty for the family.

Contents

List of Figures

List of Tables

Nomenclature

x	A visual data sample
X	Set of visual data x
N	Number of visual data in X
y	A non-visual data sample (e.g., label)
Y	Set of non-visual data y
d	Number of feature dimension
S	Source domain
T	Target domain

Major notations of deep neural network

Φ	A deep neural network (module / layer)
θ	Model parameters of Φ
$\ell(\cdot)$	A loss function
∂	Differentiation operator
n	Number of samples in each mini-batch

Major notations of k-means clustering

$\boldsymbol{\mu}$	A cluster centre feature vector
$\boldsymbol{\psi}$	A set of cluster centre features

General rules for notation definition

scalar	normal lower-case letters
set	normal UPPER-case letters
vector	**bold** lower-case letters
tensor (matrix)	**bold** UPPER-case letters

Basic Operations

$\boldsymbol{A}^T$	transpose of matrix $\boldsymbol{A}$
$tr(\boldsymbol{A})$	trace of $\boldsymbol{A}$, $\sum_i a_{ii}$
$\|\|\boldsymbol{A}\|\|_F$	Frobenius norm, $\sqrt{tr(\boldsymbol{A}^T\boldsymbol{A})}$
$\|\|\boldsymbol{A}\|\|_p$	p-norm of a vector or a matrix

Chapter 1

Introduction

Artificial Intelligence (AI), also called machine intelligence, is the intelligence demonstrated by machines in 'learning' and 'solving problems' as a human does. Benefited from the repaid developments of AI theories (e.g., deep learning and reinforcement learning) and computer techniques (e.g., Graph Process Units (GPUs) and massive storage devices), AI systems are applied in a wide range of applications such as self-driving cars, surveillance, stock management, medical diagnosis, etc. Among different data formats, visual data provides an enormous amount and the richest information. For example, self-driving cars usually equipped with multiple cameras for a better understanding of road conditions. Visual intelligence thus plays a crucial role in extracting valuable visual information via effective data analysis. One of the fundamental tasks in understanding visual data is recognition. It requires the intelligent systems should be able to precisely identify what appears or is happening in the visual input. In the case of the self-driving car, all the people and vehicles around it should be visually recognised for driving safety. However, visual recognition is a challenging task, and one of the main obstacles is the visual appearances of the same entity/instance can be different. This is caused by many factors such as lighting conditions, occlusions, deformations, camera view angles and various recording devices. Such factors are also called views. Different types of views exist and are commonplace in many realistic visual applications. Therefore, understanding the multi-view visual data has attracted great attentions from both research and industrial communities.

1.1 Multi-View Visual Data

Visual analytic problems are challenging due to the diverse visual patterns are presented in the highly-uncertain contexts. Many factors are responsible for this. However, some of these factors are only randomly occur in limited scenarios as noises. On the contrary, the factors that can be systematically described and are widely exist in the large scale visual data are more worthy of attention. Such factors are denoted as views. In this thesis, three distinctive views in visual data are studied, as shown in Figure 1.1.

Camera The visual appearance of the same entity captured from different cameras can be drastically changed. For example, the images of a person recorded by different surveillance cameras can have very different appearance characteristics due to the camera angles and the specified environments.

Modality A view can be a specified modality of the visual data. An object/entity can be represented in different visual modalities other than the normal RGB mode. For example, the near infrared images are recorded in a face recognition system to compensate the lighting-sensitive RGB ones.

Dataset The independently collected visual datasets with relevant tasks can be treated as the multi-view data where each dataset corresponds to one view, denoted as a domain. Each domain represents a distinctive visual pattern, and the sampling bias among datasets are called domain gaps. For example, the images of digital numbers collected from hand-writings (MNIST) (LeCun et al, 1998) and street view house numbers (SVHN) (Netzer et al, 2011) form two domains with apparent visual differences. However, they are used for the same task, digit categorisation.

Hybrid Views Different types of views can simultaneously appear in sophisticated visual learning settings. One example is the unsupervised domain adaptive (UDA) person re-identification (Re-ID) problem with two person Re-ID datasets are available. Both of them serve the retrieval purpose based on the visual appearance of people. However, they are built upon different surveillance networks with distinctive environments and visual patterns. Each dataset thus corresponds to a domain. Moreover, each Re-ID dataset contains the person images captured by multiple cameras in a specified surveillance network. Therefore, each domain also contains the multi-view data on its own.

Multi-view data is not limited to the visual applications and widely appears in different topics such as translation in natural language processing (NLP). Therefore, many techniques developed

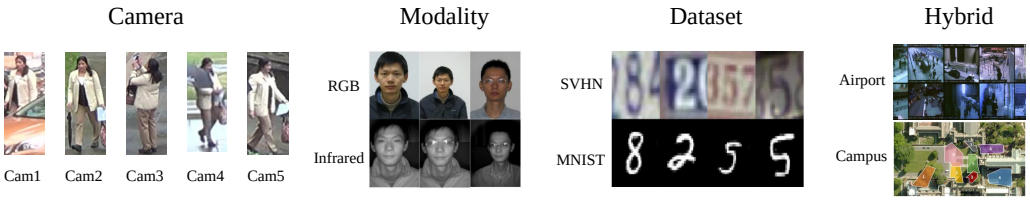


Figure 1.1: Illustrations of multi-view visual data. Different types of views are considered. Cameras: person images from different surveillance cameras. Different modalities, e.g., RGB and infrared, can be treated as specified views. Different datasets with relevant tasks as domains, e.g., two digit datasets. The hybrid case, unsupervised domain adaptive person Re-ID. Each Re-ID dataset corresponds to a specified domain. Each domain consist of multi-view data, i.e., person images from different cameras.

here can potentially benefit a variety of tasks in data mining and machine learning.

1.2 Multi-View Learning Tasks for Visual Understanding

Different learning tasks based on the multi-view visual data can be divided into three main categories, cross-view recognition, multi-view fusion and multi-view synthesis. The main characteristics of these three tasks are detailed as follows.

Cross-View Recognition The main purpose of this task is to match or retrieval the corresponding data samples of the same entity/object across different views. Many visual applications are cross-view recognition problems such as recognising the same person under different non-overlapping camera views in the person re-identification (Re-ID), as shown in Figure 1.2. The main challenge in cross-view recognition comes from the nature of multi-view data where different views contain the distinctive information of the same visual entity. To overcome this obstacle for better results, the consensus and discriminative information across different views play the key role. The consensus information bridges the gaps across views while the discriminative one helps to distinguish different entities.

Multi-View Fusion The learning objective of multi-view fusion is integrating information from different views for a more comprehensive understanding of the entity and thus improving the model performance. Taking the visual system of a self-driving car as an example. The thermal images should be captured along with the RGB ones to build a more comprehensive and robust perception on road conditions, as illustrated in the middle case of Figure 1.2. Instead of focusing on the cross-view consensus information, multi-view fusion requires the complementary and discriminative information. That is to find out the orthogonal and irreplaceable information from

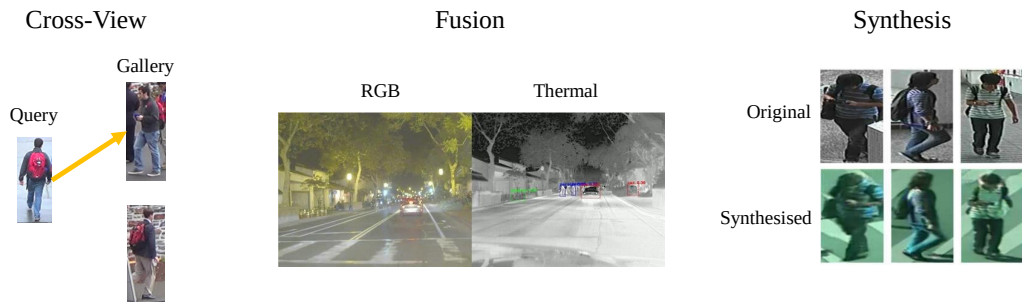


Figure 1.2: Three multi-view learning tasks, cross-view recognition, multi-view fusion and multi-view synthesis, are illustrated accordingly. Person re-identification is a typical application of cross-view recognition where the query and gallery person images are from different camera views. Fusing visual inputs from the RGB and thermal views helps the self-driving system to have a better understanding of its environment. The person images with the new styles of the other domain are generated (as shown in the second row). Nevertheless, the main characteristics such as gestures are still kept in the generated images.

each view and then fuse them for better results.

Multi-View Synthesis This final task focuses on generating the visual data samples of specified or even novel views. One obvious benefit of multi-view synthesis is to alleviate the impact of imbalance or missing data among different views. For example, in person re-identification (Re-ID), Generative Adversarial Networks (GANs) (Goodfellow et al, 2014) can be used to synthesis the person images across domains/datasets for the unsupervised domain adaptive (UDA) Re-ID problems (Deng et al, 2018; Wei et al, 2018). The generated images are adapted to the target style while keeping the main characteristics of an identity.

In this thesis, the cross-view recognition problems, which widely exist in many applications, are the main concerns. Furthermore, it is noteworthy that both cross-view recognition and multi-view fusion belong to broader multi-view recognition. Therefore, many techniques discussed and proposed here for the cross-view recognition can also be generalised to the fusion one.

1.3 Cross-View Recognition Settings

Two separated dimensions can be used to describe the different cross-view recognition settings studied in this thesis. The first dimension is the number of views, and the second dimension is the amount of annotation available for each view.

Number of Views The multi-view data consists of two views only is the most straightforward setting. However, it still attracts many attentions. Two main reasons are behind this. On the one

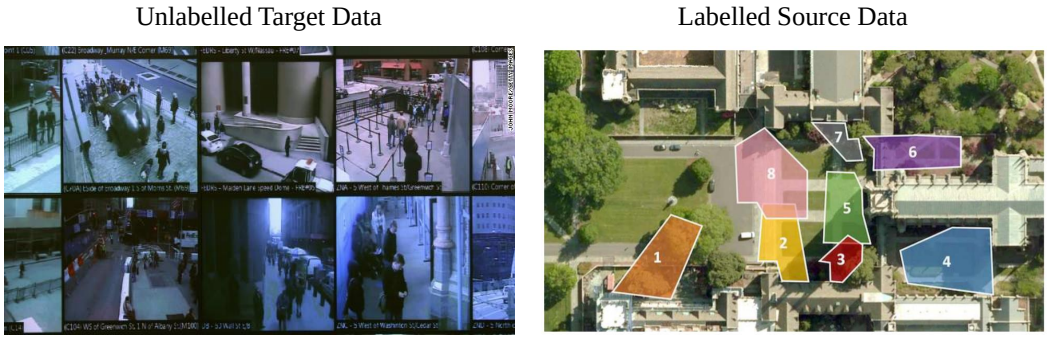


Figure 1.3: Unsupervised domain adaptive (UDA) multi-view learning in person re-identification. The target domain is from an airport surveillance. Labelling such large-scale multi-view visual data is very expensive. The other labelled Re-ID dataset (from a campus scenario) is available as source domain to improve the performance of target one.

hand, many problems follow the two-view setting. For example, a self-driving car usually uses the images from RGB and thermal modalities. On the other hand, many multi-view learning algorithms are derived from the simple two-view case before generalised to more complex situations. The more sophisticated settings are with data from more than two views. One typical scenario is that different cameras capture visual data from multiple view angles. For example, in the person re-identification problems, person images are captured by various non-overlapped cameras from distinctive view-points and environments. Moreover, as the number of views increased, the complexity of cross-view modelling can boosts drastically.

Amount of Annotations for Views In the *supervised multi-view learning*, the labels of data samples under different views are available, or the cross-view correspondences are provided as data pairs. However, exhaustively labelling or pairing all different views can be expensive, especially when different domains/datasets are considered. To improve the performance of such unlabelled dataset, another labelled dataset with relevant tasks can be exploited. It is known as the unsupervised domain adaptation (UDA) (Ganin et al, 2016) setting. Moreover, the UDA can be treated as a multi-view learning setting with the fully annotated data from one (source) domain and the unlabelled data from another (target) domain. A special UDA setting called the *unsupervised domain adaptive (UDA) multi-view learning* is the main focus here. Specifically, each domain under the UDA multi-view setting is a multi-view dataset, as the hybrid case of multi-view data described in Section 1.1. Different from the conventional UDA setting which assumes that different domains share the same label space, the UDA multi-view learning does not hold such assumption and only requires different domains have relevant multi-view learning tasks. A

concrete example is the unsupervised domain adaptive (UDA) person Re-ID, as illustrated in Figure 1.3. The source and target Re-ID datasets are captured from different surveillance networks with clear domain gaps. Both the source and target datasets serve the Re-ID purposes, but the people in the two different surveillance networks are non-overlapped and should have disjoint identity labels.

The three representative settings considered in this thesis can be coordinated with the two criteria mentioned above. The supervised two-view setting is first studied, then generalised to the multi-view learning with labels. In the challenging UDA multi-view learning, the target domain is unlabelled and learned with the labelled source domain under relevant multi-view learning tasks, but their label spaces are disjoint.

1.4 Challenges, Motivations and Solutions

To bridge the visual differences of the same entity/object under multiple views, a latent space that is inferred from the original visual space can be used to aligned them. There are different approaches to project the instances from the visual space to the latent one. The conventional paradigm makes use of handcrafted features (e.g., HOG (Freeman and Roth, 1995) and SIFT (Lowe et al, 1999)) to encode the visual information and subspace methods (Oja, 1983) for space transformations. However, the handcrafted features can hardly capture high-level visual semantics and the shallow subspace methods usually do not have enough model capacities for the sophisticated visual tasks. In this thesis, the more advanced deep neural network (DNN) methods are exploited. They provide end-to-end solutions for learning an effective latent space directly from the visual space. To align multi-view data via DNNs, the learning objectives are either explicitly optimising the deep latent space with the distance/correlation based alignment loss or implicitly forming tight clusters of multi-view data guided by the classification loss.

Challenges and Motivations Despite the promising multi-view learning results achieved by existing DNN methods, many limitations and open problems still exist and are discussed as follows.

1. **The scalability of deep CCA.** To align different views in a common space, one learning strategy is to maximise the mutual correlations across views. It is also known as canonical correlation analysis (CCA) (Hotelling, 1936; Golub and Zha, 1995). The recently proposed deep CCA models (Andrew et al, 2013; Wang et al, 2015b) aim to learn nonlinear

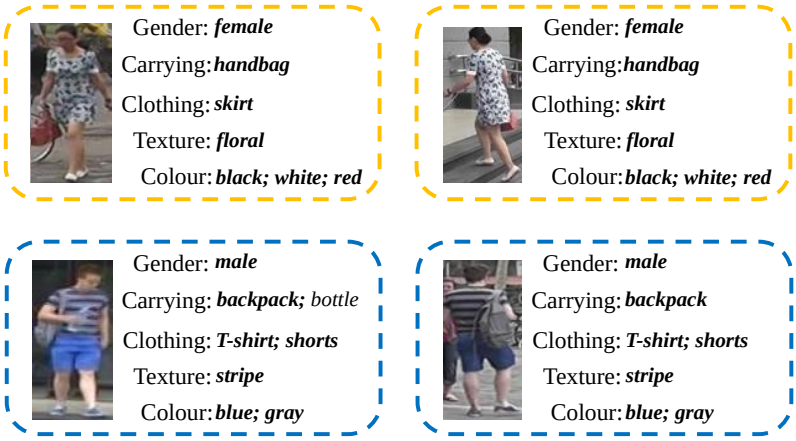


Figure 1.4: Each row shows a person captured by two camera views. A person's appearance can be described by appearance factors of multiple semantic levels for matching. Factorising the visual appearance aims to automatically discover the discriminative latent factors across views.

projections with deep neural networks rather than kernels and has been shown to be more effective than shallow CCA and KCCA. However, existing deep CCA models are inefficient and have scalability issues. The main reason is that exact or hard decorrelation is adopted. Specifically, the extracted deep feature vector for each view is decorrelated by forcing its correlation matrix over the training batch to be an identity matrix, before being minimised the distances across views in the common embedding space. Such exact decorrelation operations are computationally expensive. Either matrix inversion (Andrew et al, 2013; Wang et al, 2015b) or singular value decomposition (SVD) (Wang et al, 2015c) is required at each iteration which severely limits scalability. Furthermore, existing deep CCA models such as (Wang et al, 2015c) typically employ two separate and independent optimisation steps: the feature representation for each view is first decorrelated exactly. These decorrelation operations do not directly affect the following gradient computation and subsequent backpropagation. Without jointly optimising the decorrelation constraint and other objectives, only sub-optimal solutions are achieved.

2. **Alignments only happen in the final deep layer.** Most recent cross-view learning methods employ DNNs to learn view-invariant discriminative features. Different alignment losses are directly applied on the features from the final feature layer and the features extracted from such layer are used for matching. Such alignment mainly focuses on the high-level semantic. However, the discriminative visual semantics of multi-view data can be found at multiple levels, as illustrated in Figure 1.4. Therefore, some existing works (Ku-

mar et al, 2011; McLaughlin et al, 2017) focused on compensating the final deep feature with attribute supervisions. The application scenarios of these methods are limited by the additional visual attribute annotations, which are usually expensive to be acquired. Some other DNNs (Long et al, 2015a; Fan et al, 2018a; Yang and Ramanan, 2015) are with multi-level feature fusion architectures, but they are not designed for the multi-view learning purposes.

3. **Unsupervised multi-view dataset as target domain.** The main challenge here is the lack of annotation in the views. Considering each view corresponds to a dataset, the unlabelled dataset is denoted as the target domain. To improve the target performance, another supervised dataset is exploited as the source domain together with the unlabeled target data for training. Existing models are developed for the conventional unsupervised domain adaptation (UDA) setting with the assumption that the source and target domains are sharing the same label space. Therefore, the domain alignment framework is adopted. Different approaches differ in whether the alignment takes place in the raw image space (Hoffman et al, 2017) or a feature embedding space (Long et al, 2015b; Lin et al, 2018). However, a more challenging UDA setting, the unsupervised domain adaptive (UDA) multi-view learning, is concerned. Specifically, both domains of the UDA multi-view learning are multi-view datasets with relevant tasks. More importantly, it further assumes the label spaces across domains are disjoint, which is opposite to the conventional UDA assumption. As a result, existing domain alignment based approaches are intrinsically inappropriate for the UDA multi-view learning problems, e.g., the UDA person Re-ID.

Solutions In this thesis, different solutions are provided to the corresponding challenges mentioned above and resulting in the novel DNN based methods.

1. **Soft decorrelation in deep CCA.** To address the scalability issues in existing deep CCA methods, a robust decorrelation loss, called Stochastic Decorrelation Loss (SDL), is adopted in our proposed deep CCA, called Soft CCA. SDL is a softer constraint as the loss is only minimised rather than enforced exact decorrelation (off-diagonal elements of feature covariance matrix are all zeros). The overall learning objective in Soft CCA is consist of the decorrelation loss SDL and other losses such as the distance losses across views in the embedding space. They are all compatible with the stochastic gradient descent (SGD) in deep learning and thus jointly optimised for more global solutions.

2. **Deep factorisation and multi-level fusion.** To automatically discover the discriminative and view-invariant latent factors at multiple semantic levels for the alignment across views, a novel DNN architecture called *Multi-Level Factorisation Net (MLFN)* is proposed. The overall network is composed of multiple blocks (each of which may contain multiple convolutional layers). Each block consists of two components: A set of factor modules (FMs), each of which is a sub-network of identical architecture designed to model one factor, and a factor selection module (FSM) that dynamically selects which subset of FMs in the block are activated. Training this architecture results in FMs that specialise in processing different types of factors, and at different blocks represent factors of different semantic levels. The discovered latent factors can also be considered as latent attributes. More importantly, a compact latent semantic feature can be extracted by aggregating the FSM output vectors at all levels and enables an efficient fusion to complement the final-layer deep features.

3. **Multi-task learning for UDA multi-view learning.** Instead of following the domain alignment framework, this study presents a multi-task learning framework for the UDA multi-view learning tasks (Their comparisons are illustrated in Figure 1.5). Specifically, a shared deep feature embedding space is proposed. This space serves the source domain for supervised label prediction and the target domain with different unsupervised learning objectives. Two models are proposed based on the specified assumption made on target domain.

   The first model, termed common factorised space model (CFSM), is developed based on the idea that recognition should be performable in a shared latent factor space for both domains. In order to automatically discover such discriminative latent factors and align them for transferring knowledge across domains, our inductive bias is that input samples from both domains should generate *low-entropy* codes, i.e., near-binary codes, in this common space. This is a weaker assumption than distribution matching, but does provide a criterion that can be optimised to transfer knowledge across domains in the absence of common label space. To assist the process of the knowledge propagates down from higher-level to feature extraction for effective knowledge transfer, a novel graph Laplacian-based loss is proposed. It is built on a graph in the high-level and regularises the lower-level network feature output.

   In the second model, the unlabelled multi-view instances from target domain are utilised

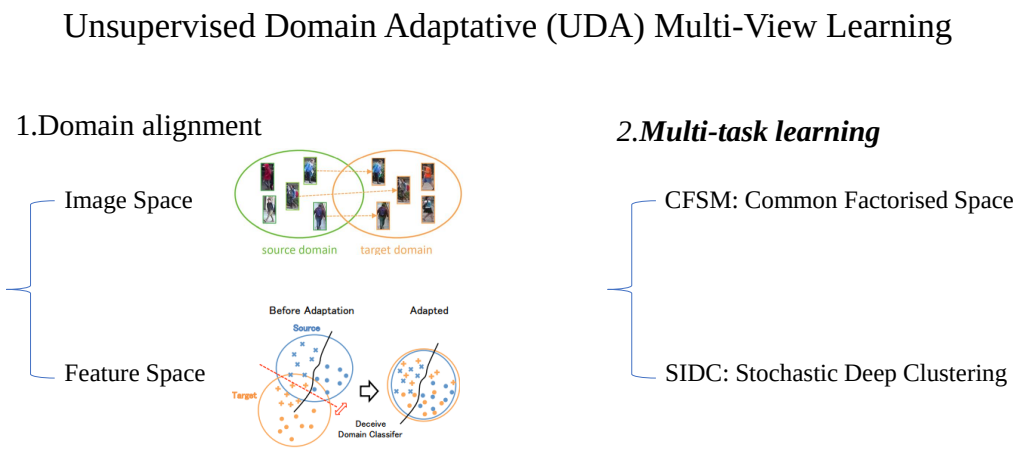


Figure 1.5: Two different pipelines for the UDA multi-view learning are illustrated. Existing work follows the domain alignment where different domains are aligned in either the image or feature representation space. In this thesis, the multi-task learning framework is adopted. Specified contributions to multi-view learning are made by the source and target domains. CFSM is built on the shared latent attribute discovery while SIDC is based on clustering the target data.

based on a simple and sound assumption: *They form clusters*, data samples of the same entity across views are potentially corresponding to a cluster. However, existing deep k-means methods are based on the hard/deterministic assignment (Xie et al, 2016a) or iterative optimisation (Yang et al, 2017a). To achieve end-to-end joint feature learning with an unsupervised clustering loss and deal with unreliable cluster assignment during training, a novel deep clustering method for the target domain called *Stochastic Inference for Deep Clustering (SIDC)* is proposed. Firstly, SIDC treats the assignment as a random variable rather than making the hard assignment of a sample to its nearest cluster. Therefore, a training sample is less likely to get stuck in a wrong cluster and it thus ameliorates the issue of reinforcing errors in deterministic approaches. Secondly, a reparameterisation trick (Jang et al, 2017) is exploited to model cluster assignment differentiably which enables end-to-end joint learning.

1.5 Contributions

The contributions made in each main chapter are summarised as follows:

1. **Chapter 3** A novel approach of deep CCA, Soft CCA, has been proposed. Comparing with existing deep CCA models, Soft CCA has two advantages. On the one hand, it is more

efficient and scalable than existing deep CCA methods - by avoiding computationally expensive operations such as SVD in existing works. On the other hand, by jointly optimising the decorrelation loss with other losses, more globally optimal solutions can be achieved. The proposed *Stochastic Decorrelation Loss (SDL)* plays the central role in our model. It formulates the decorrelation as a soft constraint and can be jointly optimised with other training objectives. Moreover, SDL is mini-batch based and the full batch statistic can be efficiently approximated by using stochastic incremental learning. While the proposed SDL is motivated by the feature decorrelation required by deep CCA learning, it can also be applied as an activation regularisation to any deep model where feature decorrelation is helpful.

2. **Chapter 4** A deep factorisation model, called Multi-Level Factorisation Net (MLFN), is proposed based on a novel deep network architecture. It makes two main contributions to multi-view learning. On the one hand, it learns to discover and dynamically identify discriminative latent factors in the visual inputs across multiple views with no attribute supervision. On the other hand, the factors computed at different levels of the network correspond to latent attributes of different semantic levels. When the selections of the factors are used as a feature and fused with the conventional deep feature, a powerful view-invariant feature representation of multi-view data is obtained. Furthermore, MLFN shows its effectiveness on general object categorisation tasks which demonstrates its potential beyond multi-view learning.

3. **Chapter 5** The multi-task learning framework is adopted for the unsupervised domain adaptive (UDA) multi-view learning problems. Two different models with distinctive assumptions made on the unsupervised target domain are proposed.

   The first method, common factorised space model (CSFM), is built on the assumption of discovering the discriminative latent attributes across both source and target domains in the shared space. Specifically, an unsupervised factorisation loss is then derived and applied on such common space which serves the purpose of optimising low-entropy criterion for discriminative latent factors discovery. On the other hand, a novel graph Laplacian based loss is proposed to better exploit the more aligned and discriminative supervision from higher-level to improve the deep feature learning. Finally, the proposed CFSM is demonstrated to be simple yet effective for different transfer learning problems other than UDA

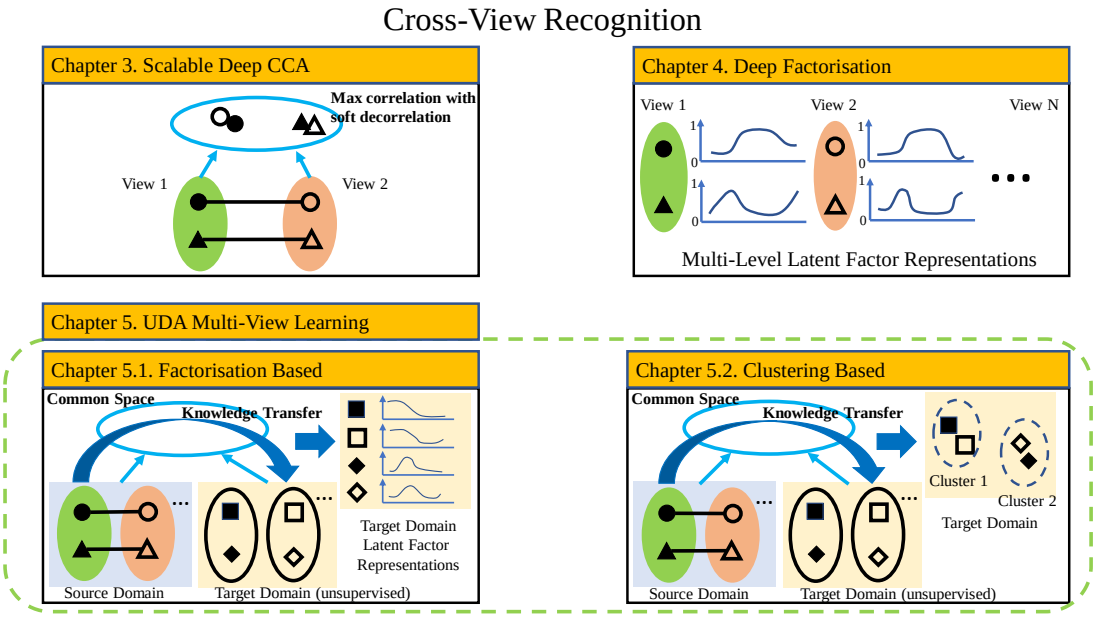


Figure 1.6: An overview of the main studies carried out in this thesis. Three important cross-view recognition problems are presented. In Chapter 3, the scalability issues of deep canonical correlation analysis (CCA) are studied with the supervised data from two views. A deep factorisation network architecture is proposed in Chapter 4 to discover the discriminative latent factors across multiple views. In Chapter 5, we focus on the more challenging unsupervised domain adaptive (UDA) multi-view learning with an unlabelled multi-view dataset as the target domain. The multi-task learning framework is adopted for this setting and two distinctive models are derived under different assumptions.

multi-view learning.

The contributions of the second model are summarised as follows. First of all, a new unsupervised domain adaptive multi-view learning model based on learning a joint feature embedding that encourages the target domain data to form clusters. Secondly, a novel deep clustering method called Stochastic Inference for Deep Clustering (SIDC) is proposed for the target domain data. It jointly learns representation and clustering by treating cluster assignment as a random, rather than deterministic, variable to alleviate compounding errors at early training. Last but not least, a new triplet loss for target data is also formulated based on the learned cluster centres and stochastic assignments to further boost the performance. Moreover, it is shown that SIDC is effective on its own on the conventional data clustering task.

To deliver a holistic understanding about this thesis, the main studies are summarised and framed into a graphical abstract as illustrated in Figure 1.6.

Chapter 2

Literature Review

Multi-View learning (MVL) is one of the fundamental tasks in machine learning where considerable effort has been spent and significant improvements have been achieved over the years (Xu et al, 2013; Zhao et al, 2017b; Li et al, 2018b). Many realistic problems can be modelled as the multi-view settings and solved by the corresponding learning algorithms (Tao et al, 2017; Guo and Xiao, 2012; Li et al, 2016). Different types of multi-view visual data are the main focus of this thesis. The common views in visual applications are camera views (Du et al, 2014; McLaughlin et al, 2017) and modalities (Li et al, 2013; Yu et al, 2016). Different datasets with relevant tasks are known as domains (Ganin and Lempitsky, 2015) and are treated as a special kind of views. In general, these MVL problems can be summarized into three categories, cross-view recognition (Lin et al, 2015b; Li et al, 2017b), multi-view fusion (Yu et al, 2013) and multi-view synthesis (Tian et al, 2018; Sun et al, 2018), as shown in Figure 1.2. In this thesis, the main concentrations are on solving cross-view visual recognition problems. Three representative settings in cross-view recognition are studied. They differ in two axes, the number of views considered and the amount of supervision provided for views. They are listed as follows, evolving from the simple to the more challenging ones,

1. Supervised multi-view data with two views only. This is the most simplified case in MVL where many MVL algorithms starts with. The data samples of the same entity/object across views are paired for model training.

2. Generalised multi-view setting. It is a more general setting with data samples from more

than two views are considered. The supervision of the entities/objectives across different views are also provided.

3. Unsupervised domain adaptive (UDA) multi-view learning. It is the most sophisticated one among the three settings. Different from the previous two settings with full supervision provided, the target domain (a multi-view dataset) is unsupervised. To improve the target performance, a labelled multi-view dataset with relevant task is provided as source domain. Different from the conventional UDA (Ganin and Lempitsky, 2015), the assumption of the aligned label spaces across domains are not hold in the UDA multi-view learning. In contrast, it assumes the disjoint label spaces of different domains.

Substantial efforts have been made towards solving the multi-view learning problems for visual understanding with different deep neural networks (DNNs). A brief overview of the significant deep learning schemes used in the visual recognition problems is provided in Section 2.1. The progress made on the three cross-view recognition settings mentioned above is then reviewed. Deep CCA combines the Canonical Correlation Analysis (CCA) objective with DNN as non-linear projection for the two-view supervised learning problems. However, existing methods have scalability issues, as detailed in Section 2.2. In Section 2.3, different advanced DNN architectures for the supervised multi-view learning are discussed. They are also compared with the proposed deep factorisation network. Finally, the connections and differences between the unsupervised domain adaptive (UDA) multi-view learning and other transfer learning settings are inspected in Section 2.4.

2.1 Visual Recognition Based on Deep Learning

2.1.1 Deep Hierarchical Architecture

In order to handle the large-scale visual data such as ImageNet (Deng et al, 2009) and surveillance data (Zheng et al, 2017), the convolution based deep neural networks (DNNs) (LeCun et al, 2015; Krizhevsky et al, 2012; He et al, 2016a) are deployed, and such deep models usually achieve better performance than the non-deep/shallow ones. As shown in Figure 2.1, the discriminative semantics in the raw image space are not structured. They are naturally ambiguous, uncertain and highly unreliable due to uncontrollable factors, e.g., cluster context and noise. The DNN provides a unified approach to process the raw visual data and handle these issues. It is widely acknowledged (LeCun et al, 2015; Szegedy et al, 2015a; Jin et al, 2016) that the power of DNN

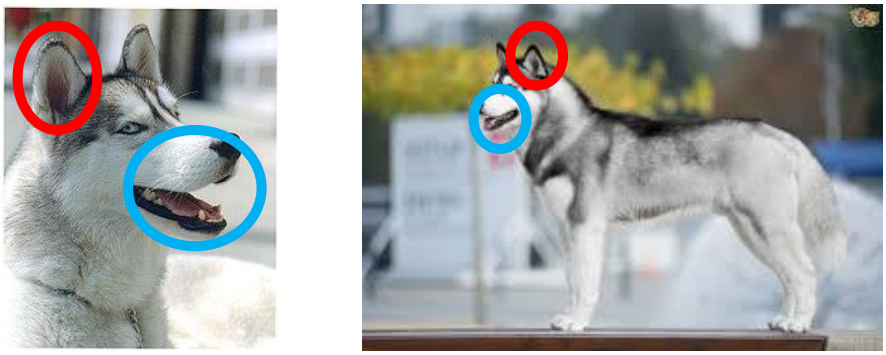


Figure 2.1: The discriminative visual semantics (highlighted with circles) in the raw images are not structured. It is hard to represent them with low-level features only since they are ambiguous and variant to scales, positions as well as view angles.

comes from the multiple levels of representation transformations via stacking non-linear modules from bottom to top to extract the more abstract and higher level of semantics in deep features. Such ability of DNN plays the key role in visual recognition problems.

2.1.2 Deep Model Optimisation

The most widely adopted DNN learning strategy is the stochastic gradient descent (SGD) (Kiwiel, 2001; Bottou, 2010). The mini-batch organisation and the mini-batch based differentiable objectives are two key components of SGD. Firstly, the full batch learning is inefficient or infeasible for DNN training with large scale data. Mini-batch input is thus a more reasonable choice by randomly sampling a small amount of instances for one step training and traversal with more iterations. Secondly, the training objectives play the crucial roles in the final performance. Specified training objectives are proposed for different tasks. Several alignment losses are widely used for deep multi-view learning. For example, Canonical Correlation Analysis (CCA) aligns different views in the correlation sense (Andrew et al, 2013). Minimising the Euclidean distance (L_2 loss) of the corresponding latent features across views can also be a simple alignment loss (Li et al, 2018b). Another widely used alignment loss is based on the positive and negative pairs in a triplet, known as the triplet loss (Chechik et al, 2010; Schroff et al, 2015; Hermans et al, 2017). It requires the distance between the anchor sample and the negative sample representations is greater than the distance between the anchor and positive representations (with a margin). The losses mentioned above explicitly align instances across views in the latent space. On the contrary, the classification loss is used for implicit alignment purpose (e.g., in person Re-ID (Zheng et al, 2016a) and face recognition (Parkhi et al, 2015)) of encouraging data samples of the same

entity (has the same labels) form a tight cluster.

2.2 Canonical Correlation Analysis

CCA is a multivariate statistical method for multi-view problems (Uurtio et al, 2018), first developed in (Hotelling, 1936). Conventional CCA is a linear model and it also generalised to the non-linear versions such as kernel CCA (KCCA) (Hardoon et al, 2004) and Deep CCA (Andrew et al, 2013). The linear and non-linear CCA objective can be unified as below following (Golub and Zha, 1995),

$$\begin{aligned} &\underset{\omega_1,\omega_2}{\operatorname{argmin}} \frac{1}{2}||P_{\omega_1}(\tilde{X}_1) - P_{\omega_2}(\tilde{X}_2)||_F^2, \\ &s.t.\ P_{\omega_1}^T(\tilde{X}_1)P_{\omega_1}(\tilde{X}_1) = P_{\omega_2}^T(\tilde{X}_2)P_{\omega_2}(\tilde{X}_2) = I, \end{aligned} \tag{2.1}$$

where $\tilde{X}_i$ represents the input data samples from view $i, i \in \{1,2\}$ with corresponding projection function $P_{\omega_i}(\cdot)$. The generalised case beyond two views will be discussed in Section 2.2.3. The model parameters of $P_{\omega_i}(\cdot)$ is $\omega_i, i \in \{1,2\}$. The covariance matrix of each view should be $I \in \mathbb{R}^{d\times d}$, i.e., the identity matrix, with d denotes the common space feature dimension. In other words, such objective can be interpreted as the feature dimensions of each view is decorrelated before different views are maximally correlated in a common space. The main differences between linear and non-linear CCA models lie in the definitions of P_{ω_i}. $P_{\omega_i}(\cdot)$ is a linear function of the input with weight matrix $\omega_i := W_i$ in CCA. The non-linear projection functions of KCCA is inexplicitly defined by the kernels (Hofmann et al, 2008) used. On the contrary, the non-linear projections in Deep CCA are explicit, the deep neural networks $P_{\omega_i} := \Phi_{\theta_i}$ and the model parameters $\omega_i := \theta_i$.

2.2.1 Deep Canonical Correlation Analysis

Inspired by the success of deep neural networks (DNNs) in representation learning (Zhou et al, 2014), Deep CCA has received increasing interest (Andrew et al, 2013; Wang et al, 2015b,c). A deep CCA architecture was first proposed by Deep CCA (DCCA) (Andrew et al, 2013) which directly computes the gradients of CCA objective and requires both a second-order optimisation method (Nocedal and Wright, 2006) and full-batch training inputs. It thus cannot cope with large training data size. An alternative deep CCA objective and architecture are proposed in Stochastic Deep CCA (SDCCA) (Wang et al, 2015c) which make it suitable for mini-batch stochastic optimisation. However, due to the exact decorrelation used, SDCCA still requires a costly SVD

operation at each training iteration. SVD's $O(d^3)$ cost is not scalable to the large layer sizes d (e.g., $d = 1024$) common in contemporary DNNs. In fact, all existing deep CCA models (Andrew et al, 2013; Wang et al, 2015b,c) take an exact decorrelation step, which limits their scalability and effectiveness as mentioned earlier. Furthermore, the exact decorrelating operations often do not directly impact the following gradient computations and backpropagation, which could lead to sub-optimal optimisation. The aforementioned issues can be alleviated or handled by formulating the decorrelation constraint as a loss which is optimised end-to-end jointly with other losses in a standard SGD procedure, making it both more scalable and more effective.

2.2.2 Decorrelation Loss

Beyond multi-view learning, many other deep models benefit from decorrelation of activations in a neural network layer. For these models, a decorrelation loss such as the proposed Sotchastic Decorrelation Loss (SDL) can be employed. Two such models are studied in this work, namely the Factorisation Autoencoder (FAE), and convolutional neural network (CNN) based classifiers. For each model, an alternative decorrelation loss exists.

FAE and XCov Loss Recently interest has regrown in models for disentangling the underlying factors of variation in the appearance of objects in images, for example identity and viewpoint (Zhu et al, 2014; Wen et al, 2016; Karaletsos et al, 2015; Veit et al, 2017; Kingma and Welling, 2014; Mathieu et al, 2016; Makhzani et al, 2015). FAE achieves semi-supervised disentangling of latent factors via a two-branch autoencoder. Recently it has been shown in (Cheung et al, 2015) that the efficacy of FAE can be improved by adding a decorrelation loss (termed XCov in (Cheung et al, 2015)) to explicitly decorrelate the computed latent factor representations. Like our SDL, computing XCov is also a mini-batch operation. But it only eliminates correlations across and not within each factor; and it computes covariance only within each mini-batch, while our SDL approximates full-batch statistics using stochastic incremental learning. SDL is demonstrated to be more effective than XCov for helping FAE to disentangle latent factors via the experiments in Sec. 3.1.4.

CNN Classifier and DeCov Loss Using CNN with a classification loss (e.g., cross entropy) for object recognition is perhaps the most popular application of deep learning in computer vision. CNN classifiers are used for not only object category recognition tasks (Krizhevsky and Hinton, 2009; Krizhevsky et al, 2012) but also object instance/identity recognition/verification tasks such as face verification (Sun et al, 2014) and person re-identification (Xiao et al, 2016).

	Purpose	Mini-batch input	Full-batch statistics	Note
Hard Decorr. (Andrew et al, 2013)	D	×	✓	2^{nd} Order Opt.
XCov (Cheung et al, 2015)	MD	✓	×	L_2 loss
BN (Ioffe and Szegedy, 2015)	R	✓	✓	STD
DeCov (Cogswell et al, 2015)	D	✓	×	L_2 loss
SDL	D	✓	✓	L_1 loss

Table 2.1: Comparisons of decorrelation losses and operations. 'D' stands for decorrelation. 'MD' denotes mutual decorrelation. 'R' is regularisation. The first method (Andrew et al, 2013) is hard decorrelation with computational expensive second order optimisation (Nocedal and Wright, 2006). Batch Normalisation (BN) is an operation for standardisation (STD) (Mendenhall and Sincich, 2016) rather than decorrelation.

When training CNNs for classification, avoiding overfitting, saturation and slow convergence are crucial (Glorot and Bengio, 2010). These problems are often alleviated by regularisation such as Batch Normalisation (BN) (Ioffe and Szegedy, 2015) and dropout (Srivastava et al, 2014). Recently it was shown that decorrelation losses can also be used for effective overfitting reduction (Cogswell et al, 2015). Compared with the existing decorrelation loss DeCov (Cogswell et al, 2015), our SDL has the following advantages: (1) More accurate covariance statistics due to full-batch approximation instead of the pure mini-batch statistics used in DeCov (Cogswell et al, 2015). (2) SDL uses a more robust L_1 formuation instead of the L_2 one in DeCov (Cogswell et al, 2015), which encourages sparser correlation and thus stronger decorrelation.

To highlight the differences and contributions, the comparisons among the different decorrelation losses mentioned above are listed in Table 2.1.

2.2.3 Multi-View Canonical Correlation Analysis

CCA models can handle the multi-view learning problems with more than two views, following the multi-view CCA (Gong et al, 2014). Assuming the total number of views is V ($V > 2$), the learning objective of multi-view CCA is similar to Eq. 2.1,

$$\begin{aligned} &\underset{\omega_1,\omega_2,\ldots,\omega_V}{\operatorname{argmin}} \frac{1}{2}\sum_{i,j=1}^{V} ||P_{\omega_i}(\tilde{X}_i) - P_{\omega_j}(\tilde{X}_j)||_F^2, \\ &s.t.\ P_{\omega_i}^T(\tilde{X}_i)P_{\omega_i}(\tilde{X}_i) = P_{\omega_j}^T(\tilde{X}_j)P_{\omega_j}(\tilde{X}_j) = I, i \neq j, \end{aligned} \tag{2.2}$$

Eq. 2.2 generalises to the multi-view case by exhaustively pairing different two views and summing up their CCA objectives. It has three main disadvantages in handling the multi-view settings ($V > 2$). (1) The number of exhaustive pairs grows fast, with the speed of $O(V^2)$. For example, there are 28 different view pairs when $V = 8$. (2) The overall model size can hardly be handled since CCA requires the view-specified modelling, i.e., $P_{\omega_i}(\cdot), i \in \{1,2,...,V\}$. It grows as V increases. The situation in Deep CCA can be even worse since V independent deep networks are deployed. (3) The training strategies and paces of different view pairs and models can be diverse. Coordinating them for the holistic optimisation in Eq. 2.2 is challenging, especially in deep CCA.

2.3 Deep Neural Network for Cross-View Recognition

Key to effective cross-view recognition is to learn the discriminative and invariant features across views for matching. Most recent works (Li et al, 2018b; Wang and Deng, 2018; Zheng et al, 2016a) are developed based on different deep neural networks (DNNs) with the advanced architectures. For example, the AlexNet (Krizhevsky et al, 2012) is used for the face recognition problem (Parkhi et al, 2015). The feature representation are extracted from the very top feature layer of a trained model. As a DNN comprises multiple feature extraction layers stacked one on the top of each other, the visual concepts of higher semantic levels are captured when forwarding from the bottom to the top layers, as pointed out in (LeCun et al, 2015; Szegedy et al, 2015a; Jin et al, 2016). It is thus infeasible for the final layer of the network to capture discriminative visual features of all semantic levels on its own. However, for cross-view recognition purposes, discriminative and view-invariant factors of multiple semantic levels should be ideally preserved in the learned features.

2.3.1 Deep Neural Networks for Person Re-ID

Specifically, a challenging cross-view recognition problem, i.e., person Re-ID (Zheng et al, 2016a; Li et al, 2017c; McLaughlin et al, 2017), is investigated. The person images in a Re-ID dataset are captured from a surveillance system with multiple non-overlapped camera views. The visual appearance of the same person across different views can change drastically due to realistic variations such as occlusion, low resolution, illumination changes and blurs (Zheng et al, 2016a; Xiao Wang, 2019).

Figure 2.2: A women in similar dressing, e.g., the red coat, but with distinctive 'looks' (Hsiao and Grauman, 2017) or 'styles' (Takagi et al, 2017) which are latent visual concepts and hard to be quantified.

Most recent person Re-ID methods train DNNs with various learning objectives including classification, verification and triplet ranking losses (Xiao et al, 2016; Sun et al, 2017; Li et al, 2017c; McLaughlin et al, 2017). Once trained, these models typically extract visual features from the final layer of a network for matching.

One approach to obtaining an appearance feature containing information from multiple semantic levels is training to predict visual attributes (Layne et al, 2012; Lin et al, 2019). By defining and annotating diverse attributes at multiple semantic levels, and training to predict them, these models are forced to encode attribute information using their top-layer features (McLaughlin et al, 2017; Schumann and Stiefelhagen, 2017; Lin et al, 2019; Khamis et al, 2014; Matsukawa and Suzuki, 2016). However, learning with the pre-defined attribute labels has two main drawbacks, and limiting the efficiency and effectiveness of corresponding models.

1. The attribute annotation process is costly and error-prone. It usually require a manual definition of the attribute dictionary and large scale image-attribute annotation, making this approach non-scalable.

2. The pre-defined attributes are semantic ('describable') but not necessarily discriminative (Farhadi et al, 2009; Xiao Wang, 2019; Hsiao and Grauman, 2017). Therefore, only sub-optimal performance can be achieved by learning with such manual attribute labels. Furthermore, some discriminative attributes are with latent semantic and hard to be clearly stated and quantified, as illustrated in Figure 2.2.

Another apporach is to complement the final layer feature with features from other layers.

A couple of studies fused representations from multiple levels (Zhao et al, 2017a; Liu et al, 2017), but this required extra effort such as body-part detection (Zhao et al, 2017a) or attention mechanisms (Liu et al, 2017) as the pre-defined auxiliary tasks.

2.3.2 Related Deep Neural Network Architectures

DNNs with Multi-Level Feature Fusion Multi-level fusion architectures have been developed in other computer vision tasks. In semantic segmentation (Long et al, 2015a; Fan et al, 2018a; Hariharan et al, 2015), feature maps from selected levels are used with shortcut connections to provide multiple granularities to the segmentation output. In visual recognition, deep features from a few selected layers were merged together to improve the final-layer representation (Yang and Ramanan, 2015; Cai et al, 2017; Yu et al, 2017c; Yang et al, 2016b). However, features extracted from limited and manually-specified layers may not reflect the optimal choice for complementing the final representation. Very few fusion architectures on specific tasks, e.g., edge detection (Xie and Tu, 2015), fuse features from all layers/levels. These models are usually designed to have limited levels (e.g., $3 \sim 5$), so their expressibility is limited.

Relevant DNN Blocks Instead of constructing each block with holistic modules as in (Krizhevsky et al, 2012; Simonyan and Zisserman, 2015; He et al, 2016a), a split-transform-merge strategy (Szegedy et al, 2016) is used to construct the modularised block architecture in ResNeXt (Xie et al, 2016b). A group of sub-network modules with duplicate structures are equally activated with their outputs summed up. Our proposed Multi-Level Factorisation Net (MLFN) leverages the ResNeXt design pattern, but extends it to include a dynamic selection of which module subset activates within each block for each image. This allows MLFN blocks to specialise in processing different latent appearance factors, and the FSM output vectors to encode a compact descriptor of detected latent factors at the corresponding level.

Our MLFN block is also related to that of Mixture-of-Expert (MoE) ones (Jacobs et al, 1991; Yuksel et al, 2012; Ahmed et al, 2016). In MoEs, a softmax activation module aims at identifying a single expert to process a given input instance. Mixture-of-Experts Layer (MoEL) methods (Eigen et al, 2014; Shazeer et al, 2017) extend flat MoE to a stacked model. They have been used to separate localisation and classification tasks in a two-level MoEL model (Eigen et al, 2014), or to implement very large neural networks by allowing each node in a cluster to run one expert in one layer of the large network (Shazeer et al, 2017). On the contrary, MLFN dynamically detects *multiple* latent factors at each level that explain each input image jointly (e.g., a person can have

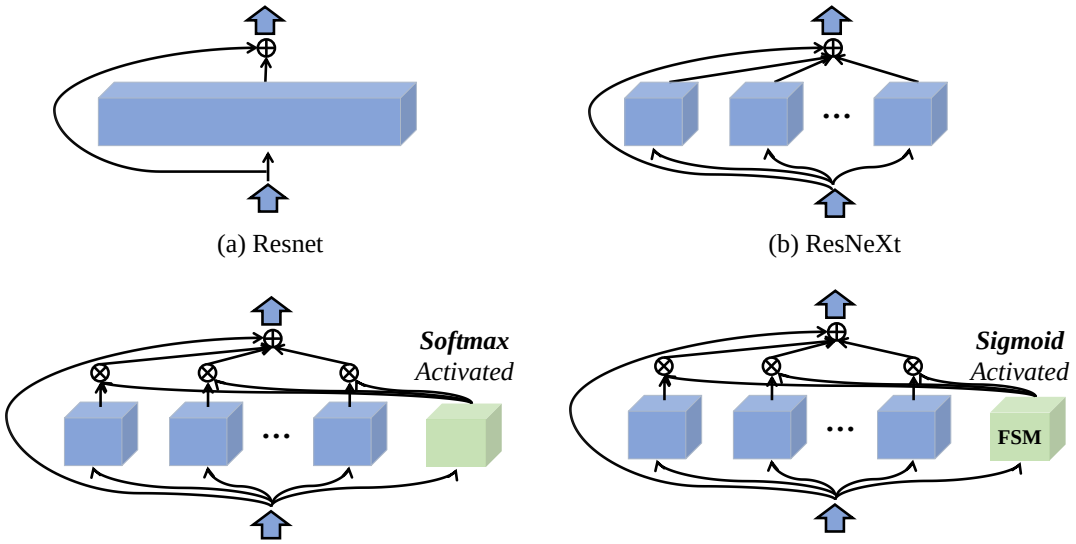


Figure 2.3: Four DNN blocks are illustrated for comparison. From (a) to (d), they are building blocks belong to Resnet (He et al, 2016a), ResNeXt (Xie et al, 2016b), Mixture-of-Experts (MoE) (Eigen et al, 2014) and the proposed MLFN. Each blue module represents an independent CNN sub-network. The longer it is, the more channels it handles. The green box is the gating sub-module to modulate the outputs of CNN sub-networks. Its outputs are softmax activated before modulation in MoE. In our MLFN, it is called factor selection module (FSM) with sigmoid activation.

both long hair and carry a bag). Thus MLFN block uses *sigmoid activated* FSM rather than *softmax* as used in MoE/MoEL, which assumes a single expert should dominate. Illustrations of the different DNN blocks, including the proposed MLFN one, are compared in Figure 2.3.

2.3.3 Other Visual Applications with Attributes

Attributes play the important role in aligning multi-view visual data, as mentioned above. The latent factors discovered by our proposed MLFN can also be interpreted as latent attributes with semantic meanings, as demonstrated in Section 4.2.4. According to (Ferrari and Zisserman, 2008), visual attributes are qualities of objects/entities from multiple perspectives such colors, textures and shapes. They are usually considered to be more descriptive than the class/category labels (Farhadi et al, 2009). Therefore, attributes are widely applicable to many other visual tasks such as object categorisation (Nagarajan and Grauman, 2018; Farhadi et al, 2009; Ferrari and Zisserman, 2008) and zero-shot learning (Lampert et al, 2009, 2013; Romera-Paredes and Torr, 2015). However, their visual attributes are mainly come from manual annotations.

2.4 Unsupervised Domain Adaptive Multi-View Learning

Labelling a large-scale visual dataset from scratch can be prohibitively expensive (Berinsky et al, 2012; Deng et al, 2009; Zheng et al, 2015). To improve the model performance on the unsupervised target dataset, one learning strategy (Zhang et al, 2019; Ganin and Lempitsky, 2015; Luo et al, 2017) is incorporating another labelled source dataset with relevant tasks for training. Each dataset is denoted as a domain and this setting is also known as unsupervised domain adaptation (UDA) (Long et al, 2015b; Ganin et al, 2016). The main purpose of UDA is to transfer knowledge from the source domain to the target one.

In this thesis, the main concentration is a specified UDA setting, the unsupervised domain adaptive (UDA) multi-view learning, where both domains are multi-view datasets with relevant tasks. Many multi-view visual applications (Zhang et al, 2019; Sohn et al, 2017) follow this setting and our main focus is the unsupervised domain adaptive (UDA) person Re-ID (Kodirov et al, 2016; Wei et al, 2018; Wang et al, 2018). The main challenges are the disjoint label (e.g., person identity) spaces across domains and the clear domain gaps, as illustrated in Figure 1.3. Moreover, the relevant transfer learning settings such as the conventional unsupervised domain adaptation (UDA) (Section 2.4.1) and the disjoint label space transfer learning (DLSTL) (Section 2.4.2) are reviewed and compared as in follows.

2.4.1 Unsupervised Domain Adaptation with Domain Alignment

An underlying assumption made by the conventional unsupervised domain adaptation (UDA) is both source and target domains share the same label space. Different domain alignment methods are thus proposed for the UDA based on this assumption. The cross-domain alignments can happen in either the raw image space (Hoffman et al, 2017; Kim et al, 2017) or a deep feature embedding space (Chen et al, 2019; Ganin et al, 2016; Gretton et al, 2009).

Although the UDA person Re-ID problem does not hold such assumption, existing methods still stick to the domain alignment framework. In (Wei et al, 2018; Deng et al, 2018), person style transfer GANs are trained to synthesise images of persons in the target domain styles, with identity information preserved from the source dataset for supervised training in the target domain. Differently, cross camera-view image synthesis in (Zhong et al, 2018) only takes place in the target domain for pseudo identity label generation. By performing joint feature learning using both domains as in our model, (Zhong et al, 2018) achieves the best results thus far. However,

it still cannot avoid the challenging GAN training process and requires the image synthesis and Re-ID models to be trained in separate stages and independently. On the other hand, domain feature alignment techniques such as maximum mean discrepancy (MMD) (Gretton et al, 2009) have been used to drive Re-ID domain adaptation (Wang et al, 2018; Lin et al, 2018). However, unlike the conventional domain-adaptation setting, the label spaces in UDA Re-ID are disjoint, so it is unclear why and how they should be aligned. Moreover, both (Wang et al, 2018; Lin et al, 2018) made use of attributes to provide a good intermediate space for alignment, but attribute annotation is not widely available, thus limiting their applicability.

2.4.2 Disjoint Label Space Transfer Learning

Transfer learning (TL) aims to transfer knowledge from one domain/task to improve performance on the other (Pan and Yang, 2009). The most widely used TL technique for deep networks is fine-tuning (Yosinski et al, 2014; Chen et al, 2018; Ren et al, 2015). Instead of training a target network from scratch, its weights are initialised by a pre-trained model from another task such as ImageNet (Deng et al, 2009) classification. Its target dataset requires to be fully supervised. Another TL setting, Disjoint Label Space Transfer Learning (DLSTL), focuses on the *disjoint label spaces* between source and target domains. The most concerned DLSTL problems are semi-superivsed DLSTL (Luo et al, 2017), i.e., both unlabelled and few labelled target data are available, and the unsupervised DLSTL, i.e., with unlabelled target data. Therefore, different UDA multi-view learning problems, e.g., the UDA person Re-ID (Zheng et al, 2016a) and fine-grained sketch based image retrieval (SBIR) (Sangkloy et al, 2016), belong to the unsupervised DLSTL. On the contrary, the conventional UDA (Ganin et al, 2016) has the same label space across domains. However, both settings have the unsupervised target datasets and the labelled source ones, as in unsupervised transfer learning (Zhang et al, 2019; Ganin and Lempitsky, 2015; Luo et al, 2017). As a summary, different transfer learning settings can find their coordinates on two axes, the relation of the label spaces across domains and the amount of target supervision provided, as illustrated in Figure 2.4.

2.4.3 Multi-Task Learning

The unsupervised domain adaptive (UDA) multi-view learning belongs to a specified transfer learning setting DLSTL, as analysed above. Therefore, the multi-task learning pipeline for transfer learning (Pan and Yang, 2009) can be adopted. Its main objective is to learn a shared deep

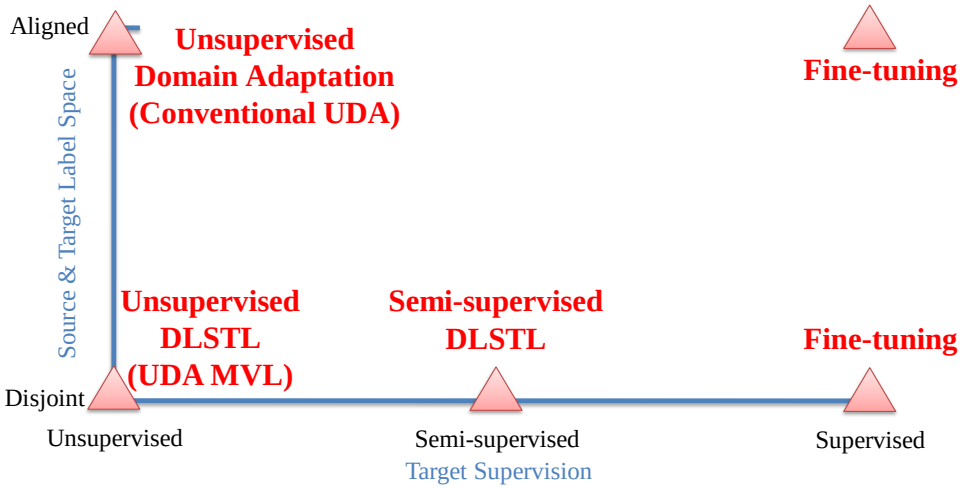


Figure 2.4: Schematic of various transfer learning problems on two criteria: the relation between source and target label space, and the amount of target problem supervision. MVL stands for Multi-View Learning.

embedding space with contributions from both domains. Specifically, the space serves the source domain for supervised learning and the target domain for unsupervised learning. The unsupervised learning objectives play the key roles in boosting the target performance. Two different assumptions on target data are proposed and results in two multi-task learning algorithms for unsupervised domain adaptive multi-view learning.

The first method is based on the *common space factorisation assumption*. Such a space is regularised to be low-entropy, i.e., near-binary. Each dimension of the space aims to capture some latent visual attributes such as colour and texture. The use of **binary codes for hashing** with deep networks goes back to (Salakhutdinov and Hinton, 2009). In computer vision, hashing layers were inserted between feature- and classification-layers to provide a hashing code (Lin et al, 2015a; Zhu et al, 2016). To produce a binary representation for fast retrieval, a threshold is applied on the sigmoid activated hashing layer (Lin et al, 2015a). **Entropy loss** for unlabelled data is another widely used regulariser (Long et al, 2016; Zhu, 2005). It is applied at the classification layer in problems where the unlabelled and labelled data share the same label space – and reflects the inductive bias that a classification boundary should not cut through the dense unlabelled data regions. Its typical use is on softmax classifier outputs where it encourages a classifier to pick a single label. **Graph-based regularisation** is popular for semi-supervised learning (SSL) which uses both labelled and unlabelled data to achieve better performance than learning with labelled data only (Zhu, 2005; Belkin et al, 2006). In SSL, graph based regularisation is applied to regularise model predictions to respect the feature-space manifold (Yue et al, 2017;

Nadler et al, 2009; Belkin et al, 2006). Moreover, exploiting the graph from lower-level to regularise higher-level features is widely adopted in other scenarios, e.g., unsupervised learning (Jia et al, 2015; Yang et al, 2017b).

The second method is based on the *clustering* assumption on the target samples. Taking person Re-ID as a example, the target instances form cluster and each cluster can be interpreted as an unknown identity. It is mostly related to **PUL** (Fan et al, 2018b), also a clustering paradigm for deep UDA person Re-ID. PUL alternates between performing k-means clustering using fixed features, and deep feature learning using fixed clusters as classification targets. Recent work shows that jointly optimising deep representations with different clustering objectives (e.g., k-means (Xie et al, 2016a; Yang et al, 2017a), agglomerative (Yang et al, 2016a), Gaussian Mixture Model (Van den Oord and Schrauwen, 2014; Jiang et al, 2017; Viroli and McLachlan, 2017), spectral (Shaham et al, 2018)) yields promising results (Aljalbout et al, 2018). Different from these clustering work, the main focus is about the source-to-target knowledge transfer with the unlabelled (target) multi-view data exploited via a clustering loss. A number of recent deep clustering methods (Xie et al, 2016a; Yang et al, 2017a) also attempt to avoid the hard/deterministic assignment in k-means clustering. **DEC** (Xie et al, 2016a) does soft cluster assignment using a Student-t distribution, and an auxiliary distribution to sharpen the initial soft assignment. Although DEC is end-to-end trainable by SGD, it fails to handle the reinforcing errors problem since the auxiliary distribution always peaks in the same position as the initial soft assignment, but with a higher probability. **DCN** (Yang et al, 2017a), on the other hand, is similar in that it also uses a reconstruction loss to regularise the clustering by avoiding trivial clustering solutions (such as mapping all the input to a single point). However, like PUL (Fan et al, 2018b), DCN is based on alternating optimisation of k-means and deep feature learning (i.e., not end-to-end trainable). It also does not address the reinforcing errors problem.

2.5 Summary

The preceding sections have discussed important studies in the literature with respect to the cross-view recognition problems in different realistic large-scale visual applications. Specifically, three representative multi-view learning settings, from the fully supervised two-view case to the more generalised case with multiple views and finally the challenging unsupervised domain adaptive (UDA) multi-view learning problems, are studied. Moreover, the deep neural network (DNN)

frameworks are exploited to handle the large-scale visual data due to their effectiveness and optimisation efficiency. As pointed out in the review, existing cross-view recognition methods still have many deficiencies. In the following chapters, novel approaches are presented to overcome the challenges as outlined below:

1. Chapter 3 **Scalable Deep Canonical Correlation Analysis:** Deep CCA generalises shallow CCA to non-linear setting via deep neural network and achieves superior performance. However, existing deep CCA models have the scalability issues due to the feature decorrelation constraints via the exact decorrelation which is computational expensive. To address these limitations, a new perspective on CCA is provided. It allows the objective to be expressed as a soften loss to be minimised by gradient descent. It results in a novel Deep CCA model, Soft CCA, that is simple to implement, more scalable and effective than existing deep CCAs. Beyond multi-view learning, the proposed Stochastic Decorrelation Loss (SDL) is applicable to a variety of tasks and models, and is superior to alternatives.

2. Chapter 4 **Deep Factorisation for Multi-View Learning:** Existing deep models exploit the features from the very top layers for cross-view recognition. Such deep features are more abstract. However, the view-invariant discriminative factors can be found at different semantic levels. To complement the final layer deep feature, existing methods rely on visual attribute annotations or inefficient fusion of outputs from multiple layers. A novel deep neural network architecture, called multi-level factorisation net (MLFN), is proposed to handle these issues. On the one hand, a novel deep factorisation building block is proposed. MLFN is built by stacking such block one by the other. It enables MLFN to discover discriminative latent factors with no additional supervision. Moreover, the task of learning multi-level factors is shared by all network blocks rather than burdening only the final layer. On the other hand, MLFN fuses information from all levels in the deep network, which is efficient because it provides a compact latent factor representation that can be easily aggregated without prohibitive feature dimension.

3. Chapter 5 **Unsupervised Domain Adaptive Multi-View Learning:** This challenging transfer learning setting consists of two multi-view datasets. One is an unsupervised target domain and the other is a fully supervised source domain. These two datasets have relevant tasks but with disjoint label spaces. For example, two person Re-ID datasets captured

from different surveillance scenarios. The main purpose of the UDA multi-view learning is to improve the target performance with the help of the labelled source domain. Existing methods follow the domain alignment framework as for the conventional unsupervised domain adaptation (UDA) and aim to align the source and target domains in either the raw image space or a deep feature embedding space. However, these methods are either hard to be trained or inappropriate for the UDA multi-view learning problems. The multi-task learning framework is adopted instead. Each domain makes specified contributions to the learning of a shared deep embedding space. Supervised learning is applied on the source domain. Two different assumptions on the unlabelled target data are proposed and result in two novel algorithms as below.

Common Space Factorisation Assumption It is built on the idea that learning should be performed in a shared latent factor space for both domains where each factor is interpreted as latent attribute. To this end, a simple yet effective model is proposed to exploit an unsupervised factorisation loss to discover a common set of discriminative latent factors between source and target datasets. And to improve feature learning for subsequent tasks such as retrieval, a novel graph-based loss is further proposed. Comprehensive experiments are conducted to show that the proposed model is effective on various transfer learning settings.

Clustering Assumption It conveys the idea that the target samples in the shared deep embedding space form clusters. Each cluster can be interpreted as an object/entity across views. Specifically, in person Re-ID, each cluster corresponds to an unknown identity of target dataset. Therefore, the key component of our proposed method is a novel Stochastic Inference for Deep Clustering (SIDC) model that encourages the unlabelled target domain to form compact clusters. Our SIDC is end-to-end trainable, and resistant to compounding errors due to the stochastic inference process. The effectiveness of SIDC is demonstrated not only on the UDA multi-view learning problems but also the general clustering tasks.

Chapter 3

Scalable Deep Canonical Correlation Analysis

Supervised cross-view recognition is a fundamental multi-view learning task with wide applications. To align different views in a joint embedding space, the correlation based learning objective is adopted and aim to maximally correlated different views for alignment. It is also known as Canonical Correlation Analysis (CCA). The recently proposed deep CCA methods aim to learn nonlinear projections with deep neural networks and has been shown to be more effective than different shallow CCA variants.

Existing deep CCA models typically first decorrelate the feature dimensions of each view before the different views are maximally correlated in a common latent space. The exact decorrelation is then applied for the feature decorrelation. Two main issues arise in these deep CCA methods because of the exact decorrelation. On the one hand, the exact decorrelation operation, either based on matrix inversion or signular value decomposition (SVD), is computationally expensive and severely limits the models scalability. On the other hand, these decorrelation operations do not directly affect the following gradient computation and subsequent backpropagation. Therefore, existing deep CCA models are optimised with separate and independent steps, which could lead to sub-optimal solutions.

A novel deep CCA approach, called Soft CCA, has been proposed in this chapter. In our model, decorrelation is formulated as a soft constraint to be jointly optimised with other training objectives. Specifically, a robust decorrelation loss, Stochastic Decorrelation Loss (SDL), is introduced, which is mini-batch based and approximates the full-batch statistics efficiently with stochastic incremental learning. SDL is a softer constraint as the loss is only minimised rather

than enforced to be zero. Therefore, Soft CCA is more efficient and scalable than existing ones. Moreover, SDL is end-to-end trainable with other losses and more globally solutions can be achieved. Last but not least, SDL can also be applied as an activation regularisation to any deep models where feature decorrelation is helpful.

The organisation of this chapter is as follows. Section 3.1 presents the details of SDL and its combinations to different deep models such as Soft CCA, Factorisation Autoencoder (FAE) and convolutional neural network (CNN). Experiments on cross-view recognition is conducted in Section 3.2. In the meanwhile, we demonstrate that SDL can be applied to a number of models for problems beyond multi-view learning. Finally, a summary is presented in Section 3.3.

3.1 Scalable Deep Canonical Correlation Analysis via Soft Decorrelation

3.1.1 Deep Canonical Correlation Analysis

Deep Canonical Correlation Analysis (CCA) extends linear CCA model by projecting views of the same entity (here images of the same objects is considered) from different views to a common latent space using a DNN with multiple branches, each corresponding to one view (see Figure 3.1). A two-view case for correlation learning is considered. Assuming the training visual data from both views are denoted as $\bar{X} = \{\bar{X}_1, \bar{X}_2\}$ where the data samples from different views are denoted as $\bar{X}_1$ and $\bar{X}_2$ respectively. The data sample correspondences across different views are given in the supervised setting. Therefore, $\bar{X}_1$ and $\bar{X}_2$ are constructed by pairing the corresponding data samples of the same objects. The number of different objects in $\bar{X}$ is denoted as N and N training pairs are thus available. The mini-batch inputs (X_1 and X_2) for deep neural network (DNN) can be organised in the similar way, by stochastically sampling n pairs from N and fed into the corresponding DNN branches. The DNN branches aim to learn functions that project paired input images into a shared latent space where they are maximally correlated. Denote the DNN projection function for view i, $i = \{1,2\}$ as $\Phi_{\theta_i} : X_i \rightarrow \boldsymbol{Z}_i$, or $\Phi_{\theta_i}(X_i) = \boldsymbol{Z}_i$ where $\boldsymbol{Z}_i \in \mathbb{R}^{n \times d}$ is the projected feature matrix for n data items for view i in the d dimension CCA embedding space and θ_i are the DNN parameters.

Following (Golub and Zha, 1995), CCA can be formulated in multiple ways and the most relevant one here is:

$$\begin{aligned} &\arg\max_{\theta_1,\theta_2} tr(\Phi_{\theta_1}^T(X_1)\Phi_{\theta_2}(X_2)), \\ s.t.\ &\Phi_{\theta_1}^T(X_1)\Phi_{\theta_1}(X_1) = \Phi_{\theta_2}^T(X_2)\Phi_{\theta_2}(X_2) = I, \end{aligned} \tag{3.1}$$

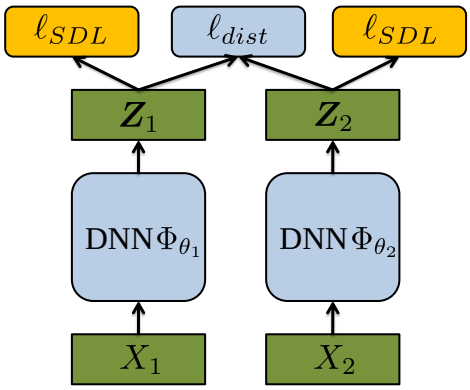


Figure 3.1: Schematic of implementing Soft CCA with SDL.

where I indicates the identity matrix. The constraints enforce exact decorrelation within each of the two input signals. Eq. 3.1 can be written into an equivalent form:

$$\begin{aligned} &\arg\min_{\theta_1,\theta_2} \frac{1}{2}||\Phi_{\theta_1}(X_1) - \Phi_{\theta_2}(X_2)||_F^2, \\ &s.t.\ \Phi_{\theta_1}^T(X_1)\Phi_{\theta_1}(X_1) = \Phi_{\theta_2}^T(X_2)\Phi_{\theta_2}(X_2) = I, \end{aligned} \tag{3.2}$$

It shows that the goal of maximising correlation between $\Phi_{\theta_1}(X_1)$ and $\Phi_{\theta_2}(X_2)$ can be achieved by minimising the L_2 distance between the decorrelated signals.

The key idea of our approach is to convert the hard constraint in Eq. 3.2 into a soft decorrelation loss to be optimised by SGD.

3.1.2 Stochastic Decorrelation Loss

The representations from one branch of a deep CCA network over a mini-batch is denoted as $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$, where n is the mini-batch size and d indicates the number of neurons/feature channels. We further assured that $\boldsymbol{Z}$ has been mini-batch standardised, i.e., each activation over the mini-batch has zero mean and unit variance. This can be easily achieved by adding a Batch Normalisation (BN) (Ioffe and Szegedy, 2015) layer.

The mini-batch covariance matrix $\boldsymbol{C}_{mini}^t$ for the t-th training step then is given as:

$$\boldsymbol{C}_{mini}^t = \frac{1}{n-1}\boldsymbol{Z}^T\boldsymbol{Z}. \tag{3.3}$$

However, full-batch statistics are required by CCA objective for decorrelation. Therefore, the full-batch covariance matrix $\boldsymbol{C}_{full}$ is approximated by accumulating statistics collected from each mini-batch. This is achieved by stochastic incremental learning. More specifically, an

accumulative covariance matrix is first computed:

$$\boldsymbol{C}^t_{accu} = \alpha \boldsymbol{C}^{t-1}_{accu} + \boldsymbol{C}^t_{mini}, \tag{3.4}$$

where $\alpha \in [0,1)$ is a forgetting/decay rate and $\boldsymbol{C}^0_{accu}$ is initialised with an all-zero matrix. A normalising factor is also computed accumulatively as $c^t = \alpha c^{t-1} + 1$ ($c^0 = 0$ initially). The final full-batch covariance matrix approximation is then computed as:

$$\boldsymbol{C}^t_{appx} = \frac{\boldsymbol{C}^t_{accu}}{c^t}. \tag{3.5}$$

If an exact decorrelation strategy as in (Andrew et al, 2013; Wang et al, 2015b,c) is followed, the off-diagonal elements of $\boldsymbol{C}^t_{appx}$ are forced to be zeros. However, that has implications on the computational cost and scalability which will be detailed later. Instead, we follow a soft decorrelation procedure and formulate the decorrelation constraint as a loss. Specifically, SDL is an L_1 loss on the off-diagonal element of $\boldsymbol{C}^t_{appx}$:

$$\ell_{SDL} = \sum_{i=1}^{d} \sum_{j \neq i}^{d} |\phi^t_{ij}|, \tag{3.6}$$

where ϕ^t_{ij} is the element in $\boldsymbol{C}^t_{appx}$ at (i,j). L_1 loss is used here to encourage sparsity in the off-diagonal elements. SDL is soft because it only penalises the correlation across activations instead of enforcing exact decorrelation. It will be jointly optimised with any other losses the model may have.

Gradients and Optimisation The gradient of ℓ_{SDL} w.r.t. z_{mi} (the element in Z at (m,i)) can be computed as

$$\begin{aligned} \frac{\partial \ell_{SDL}}{\partial z_{mi}} &= \frac{1}{c^t} \frac{1}{n-1} \sum_{j}^{k} \boldsymbol{S}(i,j) z_{mj}, \\ \boldsymbol{S}(i,j) &= \begin{cases} 1, \ \phi^t_{ij} > 0 \\ 0, \ i = j \ or \ \phi^t_{ij} = 0 \\ -1, \ \phi^t_{ij} < 0 \end{cases} \end{aligned} \tag{3.7}$$

with the sign matrix $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ and $i, j = 1, ..., d$. Eq. 3.7 can be written in a matrix form:

$$\frac{\partial \ell_{SDL}}{\partial \boldsymbol{Z}} = \frac{1}{c^t} \frac{1}{n-1} \boldsymbol{Z} \cdot \boldsymbol{S}, \tag{3.8}$$

where $\cdot$ indicates matrix multiplication.

Once the SDL gradients are computed, they are passed through the network during back-propagation and optimised along with other losses in end-to-end training.

Computational Complexity Eq. 3.6 shows that to compute the SDL in a forward pass, matrix multiplication (as in Eq. 3.3) is needed, matrix addition (as in Eq. 3.4) and matrix element-wise summation (as in Eq. 3.6). Therefore, the forward pass computation complexity of SDL is $O(nd^2)$. The gradient computation during the backward pass is in Eq. 3.8. It is also a matrix multiplication and therefore the complexity is $O(nd^2)$. The overall computational complexity of one training iteration is thus $O(nd^2)$. In contrast, existing exact decorrelation computation (Andrew et al, 2013; Wang et al, 2015c) has a complexity of $O(nd^2+d^3)$ due to the use of SVD. Note that in large scale vision problems, the number of activations in an FC layer can easily be thousands, meaning that the alternative hard decorrelation models are prohibitively expensive.

3.1.3 Stochastic Decorrelation Loss for Soft Canonical Correlation Analysis

With the proposed SDL, the constrained optimisation problem in Eq. 3.2 can be reformulated as the following unconstrained objective:

$$\arg\min_{\theta_1,\theta_2} \ell_{dist}(\Phi_{\theta_1}(X_1),\Phi_{\theta_2}(X_2)) + \lambda(\ell_{SDL}(\Phi_{\theta_1}(X_1)) + \ell_{SDL}(\Phi_{\theta_2}(X_2))), \tag{3.9}$$

where $\ell_{dist}(\Phi_{\theta_1}(X_1),\Phi_{\theta_2}(X_2))$ is the L_2 distance and λ weights the alignment versus decorrelation losses. The Soft CCA architecture is also illustrated in Figure 3.1. Note that both SDL and L_2 loss are mini-batch based losses. Therefore, Soft CCA (deep CCA model with SDL) can be realised using standard SGD optimisation for end-to-end learning.

3.1.4 Applications of Stochastic Decorrelation Loss to Other Deep Models

Factorisation Auto-Encoder (FAE) with Stochastic Decorrelation Loss (SDL) A two-factor case is described although the model generalises to an arbitrary number of factors. The two-factor FAE model is illustrated in Figure 3.2. Its encoder (a deep neural network) takes image x as input and projects it into an embedding space/latent code which has two parts: $\boldsymbol{y}$ and $\boldsymbol{z}$. Taking the digit (0-9) images of MNIST dataset as an example, $\boldsymbol{y}$ is a factor that is annotated in the training data, e.g., digit label. The other unspecified factors, such as hand writing style, are captured by $\boldsymbol{z}$. Both $\boldsymbol{y}$ and $\boldsymbol{z}$ are used as input to the decoder (e.g., a deconvolutional network) which produces a reconstruction of x, denoted as $\hat{x}$. The goal is not only to accurately reconstruct the input x, but also to represent distinct factors of variation in $\boldsymbol{y}$ and $\boldsymbol{z}$, as demonstrated in

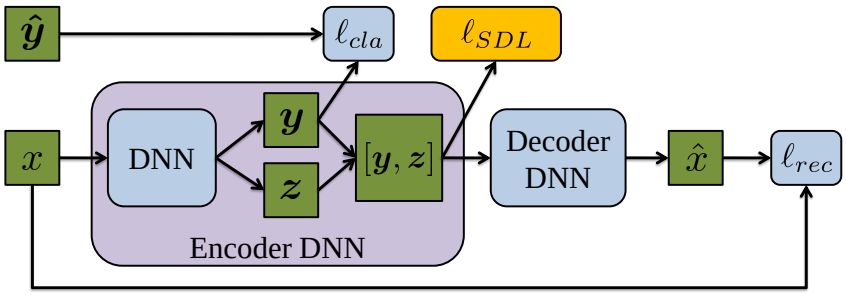


Figure 3.2: Architecture of Factorisation Auto-Encoder (FAE) with SDL.

Section 3.2.2. In statistics, such distinct representations can be described as mutual decorrelated and a decorrelation loss is thus required between $\boldsymbol{y}$ and $\boldsymbol{z}$.

Assume the FAE model is parameterised by θ. Given a training set D containing images X and their labels $\hat{Y}$ for the known factor, the learning objective of FAE is:

$$\arg\min_{\theta} L_{rec}(X,\hat{X}) + \lambda L_{cla}(Y,\hat{Y}), \tag{3.10}$$

where $L_{rec}(X,\hat{X})$ is the reconstruction loss, which is a pixel L_2 loss here, and $L_{cla}(Y,\hat{Y})$ is the classification loss, i.e., cross-entropy loss here. If there is no constraint on the relation between $\boldsymbol{y}$ and $\boldsymbol{z}$, they would not necessarily represent distinct aspects of the input signal. To disentangle them, the SDL is introduced to the objective:

$$\arg\min_{\theta} L_{rec}(X,\hat{X}) + \lambda_1 L_{cla}(Y,\hat{Y}) + \lambda_2 L_{SDL}([Y,Z]). \tag{3.11}$$

As shown in Figure 3.2, this means we decorrelate the elements of the concatenated code $[\boldsymbol{y},\boldsymbol{z}]$ which decorrelates the two code parts (factors), as well as the signal within the factors.

Convolutional Neural Network (CNN) Classifier with SDL Since decorrelation loss encourages a layer's activations to be decorrelated, it reduces activation co-adaptation and maximises the model's capacity. Therefore, SDL can be applied to each layer of a CNN classifier to boost the model performance. In our experiments, SDL is added to different CNN classifiers for different recognition tasks to demonstrate its general applicability.

3.2 Results and Analysis

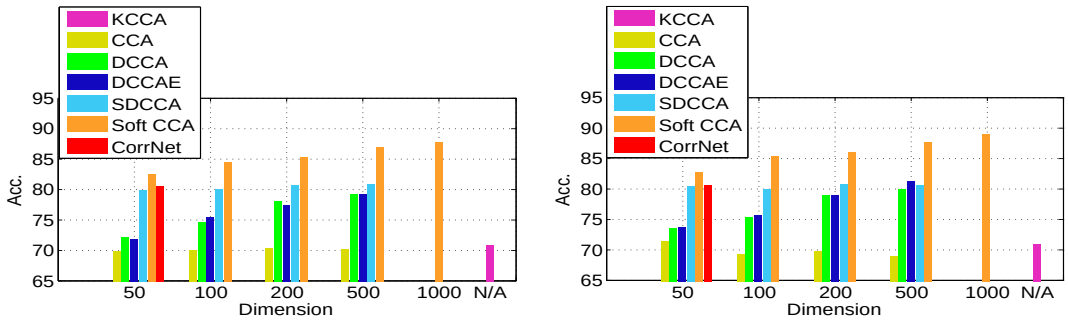
3.2.1 Soft Canonical Correlation Analysis

Datasets and Settings We evaluate the proposed Soft Canonical Correlation Analysis (CCA) and alternative deep CCA models on two widely used datasets. **MNIST** (LeCun et al, 1998)

consists of handwritten digit images with an image size of 28×28. It contains 60,000 training and 10,000 testing images respectively. The experimental setting in (Chandar et al, 2016) for cross-view recognition is followed. Deep CCA models are trained on the left and right halves of a 10,000 sized subset of training images and 5-fold cross validation is done on the provided test set for recognition. **Multi-PIE** (Gross et al, 2007) is a face dataset composed of 750,000 images of 337 people with various factors contributing to appearance variation including viewpoint, illumination and facial expression. A subset containing 6,200 images of all 337 identities in neutral expression and lighting is used. Constructing an analogous experiment to the cross-view recognition benchmark, these images are separated into the left and right view groups according to their viewing angle. Left-right view angle pairs are then formed exhaustively for the same identities to train the deep CCA models. Half of the images in both views are used for deep CCA training and a 5-fold cross validation for recognition on the rest of the data is also done.

Implementation Details For MNIST cross-view recognition, the network architecture of each view branch is identical to that in (Chandar et al, 2016) for fair comparison. Concretely, there are three hidden layer containing 500, 300, d units/activations respectively, where the d units are used as the common representation (CCA embedding layer). ReLU is applied on the hidden layers' activations (except the embedding layer). Once the CCA model is trained, on the test set, features from one view (e.g., right) are extracted, embedded with deep CCA, and then fed to a Linear SVM (Chang and Lin, 2011) classifier which is trained to recognise the images. Finally, the model is evaluated based on features from the other view (e.g., left) being projected into the shared embedding space, and recognised by the SVM. Clearly, the performance of the SVM on this cross-view recognition task depends on the efficacy of the CCA embedding. An analogous cross-view recognition setting is used for the Multi-PIE dataset. The DNN architecture for Multi-PIE also has three hidden layers: 1024, 512, d units, the d units are used as the CCA embedding layer. ReLU is applied on the hidden layers' activations (except the embedding layer).

Competitors For shallow CCA, we compare the standard linear CCA (Hotelling, 1936) and its nonlinear kernelised variant, KCCA (Hardoon et al, 2004). The KCCA results are obtained from (Chandar et al, 2016). For the deep CCA models, CorrNet (Chandar et al, 2016), DCCA (Andrew et al, 2013), DCCAE (Wang et al, 2015a) and SDCCA (Wang et al, 2015c) are compared. CorrNet (Chandar et al, 2016) combines correlation maximisation with cross-view autoencoder loss and uses Batch Normalisation. Without access to their code, only the reported result in



(a) Left-to-right

(b) Right-to-left

Figure 3.3: Cross-view digit recognition results on MNIST. Note that CCA is not scalable to a common space dimension that is greater than the total dimension of 784. Moreover, DCCA, DCCAE and SDCCA are also intractable with our GPU resources when the common space dimension becomes 1000.

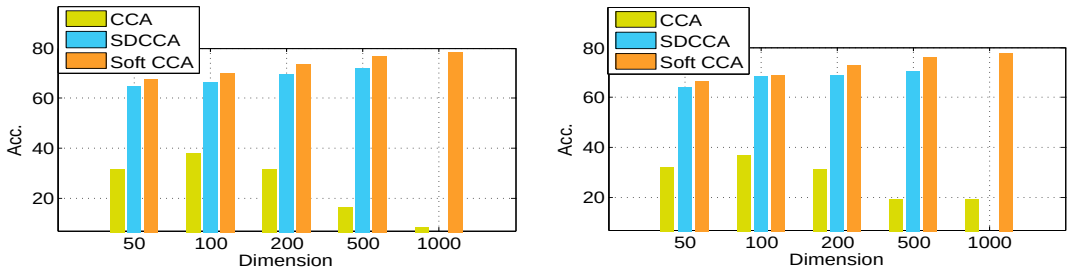
(Chandar et al, 2016), which was obtained only on MNIST with $d = 50$, is used. As far as we know, SDCCA (Wang et al, 2015c) is the most efficient state-of-the-art deep CCA model to date.

Results on Cross-View Recognition Figures 3.3 and 3.4 show the results for cross-view digit and face recognition. The following observations are made: (1) The deep models achieve better performance than the shallow ones. (2) Our Soft CCA achieves the best results on both datasets with all CCA space dimensions. (3) Increasing the common space dimension d benefits SDCCA very little and even harms the performance of other competitors (e.g. CCA). In contrast, our Soft CCA clearly benefits from larger CCA space dimensions.

	50D	100D	200D	500D	1000D
Upper Bound	50	100	200	500	1000
CCA	28.3	34.2	48.7	74.0	-
DCCA	29.5	44.9	59.0	84.7	-
DCCAE	29.3	44.2	58.1	84.4	-
SDCCA	**46.4**	**89.5**	166.1	307.4	-
Soft CCA	45.5	87.0	**166.3**	**356.8**	**437.7**

Table 3.1: Correlation strength on MNIST. '-' indicates that the result is not obtainable due to the corresponding model being intractable with our available hardware.

Results on Cross-View Correlation Another way to evaluate CCA models is to measure the average correlation strength of each matching pair of data when they are projected into the common CCA space (Wang et al, 2015c). We follow the experimental setting and network ar-



(a) Left-to-right (b) Right-to-left

Figure 3.4: Cross-view face recognition results on Multi-PIE. Accuracy (%). Note that SDCCA is intractable with our GPU resources when the common space dimension becomes 1000.

chitecture of (Wang et al, 2015c) (SDCCA) for a fair comparison. The results of MNIST and Multi-PIE are shown in Table 3.1 and Table 3.2 respectively. Following conclusions are made from the results: (1) Again the deep models achieve higher correlation values indicating that they align the two views much better than the linear CCA model. (2) For the easier digit classification task in MNIST, our model is slightly inferior to SDCCA at 50D and 100D, but better after 200D. For the more challenging face recognition problem in Multi-PIE, Soft CCA consistently outperforms SDCCA and the gap increases with the dimension. These results suggest that our model is more effective with higher dimensional embedding space, which is required for more challenging computer vision tasks.

	50D	100D	200D	500D	1000D
Upper Bound	50	100	200	500	1000
CCA	12.8	23.9	53.4	140.6	207.1
SDCCA	25.7	51.5	151.2	228.3	-
Soft CCA	**29.2**	**60.5**	**163.2**	**257.7**	**283.9**

Table 3.2: Correlation strength on Multi-PIE.

Evaluation on Scalability The training time for our model and that for the most efficient deep CCA model proposed to date, SDCCA (Wang et al, 2015c), are compared. Figure 3.5 shows that our soft CCA is always more efficient than SDCCA even at the low dimensions[1]. Importantly, when the CCA embedding space dimension approaches 4,000 (roughly the same as the final FC layer size of popular DNNs like AlexNet and VGGNet), our model is clearly much more efficient

[1]The speedup is significant even under low dimensions; it is just not very salient in Figure 3.5 due to the scaling problem. E.g, at 50D and 100D, Soft CCA is 2 and 5 time faster to train respectively.

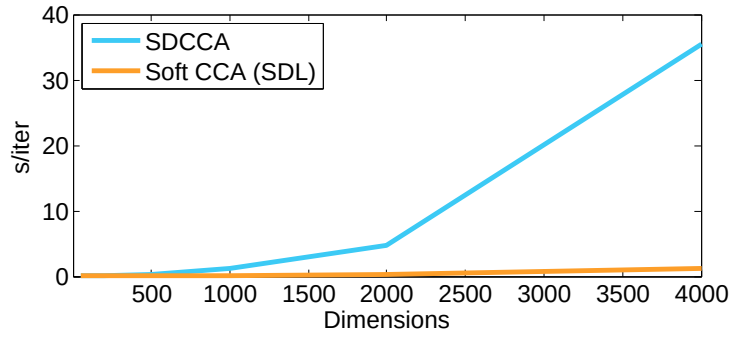


Figure 3.5: Comparing training time (seconds/iteration) on MNIST given different CCA space dimensions.

to train. This is due to the $O(d^2)$ vs. $O(d^3)$ computational complexity difference.

3.2.2 Factorisation Auto-Encoder with Stochastic Decorrelation Loss

Dataset and Settings MNIST (LeCun et al, 1998) is used, and the same experimental setting as (Cheung et al, 2015) is followed. The network architecture is 784-1000-1000-{$\boldsymbol{y}$+$\boldsymbol{z}$}-1000-1000-784, where 784 is the dimension of the vectorised image. ReLU is applied on the hidden layers' activations (except $\boldsymbol{y}$, $\boldsymbol{z}$). As shown in Figure 3.2, among the two factors to be disentangled, $\boldsymbol{y}$ is the digit class which is annotated with the training data. The other factor $\boldsymbol{z}$ corresponds to aspects of appearance besides class – i.e., the unannotated writing style. In our experiments, the dimension of $\boldsymbol{y}$ is fixed to 10 corresponding to the 10 digit classes and the dimension of $\boldsymbol{z}$ is also set to 10. The performance of a vanilla FAE (basic network with only reconstruction and classification loss), FAE+XCov (Cheung et al, 2015), FAE+DeCov (Cogswell et al, 2015) and our FAE+SDL are compared.

Evaluation on Disentanglement In the ideal case, the two factors will be completed disentangled in $\boldsymbol{y}$ and $\boldsymbol{z}$, i.e., $\boldsymbol{y}$ contains no information about the style and $\boldsymbol{z}$ contains nothing about the class. To quantify this, the digit classification performance are compared with the inferred $\boldsymbol{y}$ and $\boldsymbol{z}$ on the test set. Classification based on $\boldsymbol{y}$ is given by the prediction scores from the FAE classification branch. The inferred $\boldsymbol{z}$ requires an additional classification model and a linear SVM is trained using $\boldsymbol{z}$ from the training set and test it on the test set. Predictions based on $\boldsymbol{y}$ and $\boldsymbol{z}$ should thus ideally give perfect and random chance accuracies respectively. Table 3.3 shows that with SDL, the style feature $\boldsymbol{z}$'s classification performance is close to random guess (10%), and better (closer to random) than that of XCov and DeCov, whilst using with the vanilla FAE with no decorrelation loss, it still contains extensive class information. Meanwhile, the disentangled

	FAE	XCov	DeCov	SDL
z ($\downarrow$)	43.44	14.51	15.42	**11.35**
y ($\uparrow$)	97.23	95.72	97.09	**97.33**

Table 3.3: Disentanglement efficacy. Classification accuracy (%) using representation of each branch in MNIST FAE.

$\boldsymbol{y}$ provides the highest classification accuracy using our FAE+SDL. The results suggest that our model is more effective than the alternative XCov and DeCov in disentangling latent factors. This is because our SDL does a stochastic approximation of the full-batch statistics, whilst both XCov and DeCov only use information from each mini-batch.

Qualitative Results With the style factor disentangled from the class factor, the FAE can be used to transfer styles to a new digit. Given an input image containing a certain digit with certain handwriting style, we can keep the inferred $\boldsymbol{z}$ and change the value $\boldsymbol{y}$ manually to a different digit class. After feeding both the original $\boldsymbol{z}$ and the modified $\boldsymbol{y}$ to the decoder, a new digit can be synthesised with the same style as the input image. Qualitative results are shown in Figure 3.6. We see the better disentanglement efficacy of our model in terms of clearer digit reconstruction with clearer style transfer.

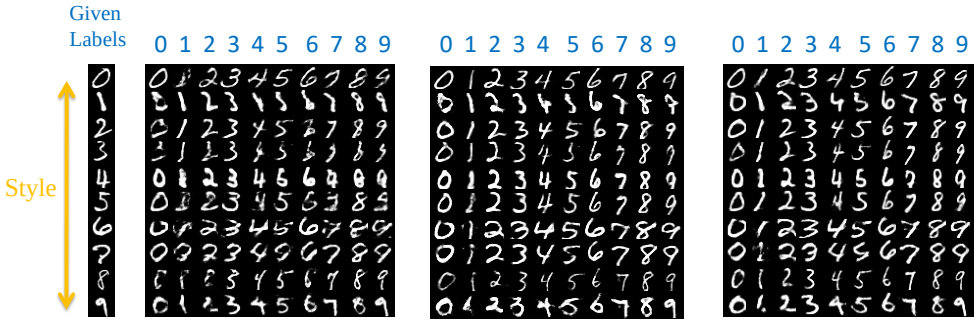


Figure 3.6: Qualitative results of handwriting style transfer with different FAE models. (a) FAE; (b) FAE + XCov (Cheung et al, 2015) ; (c) FAE + SDL. The dimension of $\boldsymbol{z}$ is set to 10.

3.2.3 Deep Classifier with Stochastic Decorrelation Loss

Experiments on Object Recognition CIFAR10 (Krizhevsky and Hinton, 2009) which consists of 60,000 32×32 colour images in 10 categories, with 6000 images per category is used. The standard experimental setting in (Krizhevsky and Hinton, 2009) is followed. The DNN

baseline model used is a 20-layer ResNet (He et al, 2016a). SDL is compared with existing decorrelation loss DeCov (Cogswell et al, 2015) and the baseline (with BN but without any decorrelation loss) in Table 3.4. The proposed SDL leads to a 1.32% performance improvement over the baseline model and also outperforms the alternative DeCov loss by 0.82%.

	Baseline	DeCov	SDL
Accuracy	91.12	91.62	92.44

Table 3.4: CIFAR10 classification results (%)

Person Re-Identification In this experiment, a CNN classifier is applied to solve a more challenging recognition problem. The person re-identification (Re-ID) problem aims to match pedestrians captured by non-overlapping CCTV cameras[2]. One of the biggest and most popular Re-ID benchmarks is used. **Market-1501** (Zheng et al, 2015) is collected from 6 different cameras. It has 32,668 bounding boxes of 1,501 identities obtained using a Deformable Part Model (DPM) person detector. Following the standards split (Zheng et al, 2015), we use 751 identities with 12,936 images for training and the rest 750 identities with 19,732 images for testing. Experiments are conducted under both the single-query and multi-query evaluation settings[3]. The Rank-1 accuracy is computed to evaluate all the methods. The mean average precision (mAP) (Zheng et al, 2015) is also calculated. For the base model, we use one of the state-of-the-art deep Re-ID models, DGDNet (Xiao et al, 2016), which is built on Inception modules (Szegedy et al, 2015b). Our model (DGDNet+SDL) adds SDL on the output of each BN layer in DGDNet during training.

The results are shown in Table 3.5, along with some high performing state-of-the-art alternatives. We can see that: (1) Our model (DGDNet+SDL) outperforms a number of state-of-the-art alternatives. (2) Compared to the base model (DGDNet without decorrelation loss), adding our SDL boosts the performance by a clear margin. (3) When the alternative DeCov loss is added to the base model, its performance is also improved, but by a smaller margin. This result thus

[2]Note that person Re-ID is a multi-view learning problem with multiple (more than two) views. Deep CCA is inefficient in handling this, as mentioned in Sec. 2.2.3. The state-of-the-art Re-ID approaches train a holistic deep model with identity-supervised classification loss.

[3]In Person Re-ID, multiple query (M-Query/MQ) is an evaluation setting assumes a person has multiple query images that are under the same camera view. Therefore, the query feature of that person is computed by averaging or max pooling the features of such query images. In contrast, single query (S-Query/SQ) mode has one query image only. Comparing with the single query results, multiple query ones are usually better since the multiple query feature contains more information. However, single query is the default evaluation method across all datasets.

	S-Query		M-Query	
	mAP	R1	mAP	R1
Siamese LSTM (Varior et al, 2016b)	–	–	35.3	61.6
Gated S-CNN (Varior et al, 2016a)	39.55	65.88	48.45	76.04
CNN Embedding (Zheng et al, 2016b)	59.87	79.51	70.33	85.84
Spindle (Zhao et al, 2017a)	-	76.9	-	-
HP-net (Liu et al, 2017)	-	76.9	-	-
OIM (Xiao et al, 2017)	-	82.1	-	-
Re-rank (Zhong et al, 2017)	63.6	77.1	-	-
DPA (Zhao et al, 2017c)	63.4	81.0	-	-
SVDNet (Sun et al, 2017)	62.1	82.3	-	-
Context (Li et al, 2017a)	57.5	80.3	66.7	86.8
JLML (Li et al, 2017c)	64.4	83.9	74.5	89.7
DGDNet*	64.55	85.06	73.30	89.40
DGDNet+DeCov (Cogswell et al, 2015)	65.74	85.86	74.72	90.53
DGDNet+SDL	**67.67**	**86.75**	75.77	**91.06**

Table 3.5: Market-1501 Results. S-Query means Single Query, and M-Query means Multiple Query. '–' indicates no reported result. DGDNet* refers to the basic network used in DGD (Xiao et al, 2016), but trained from scratch only on Market-1501, without multi-task learning through the Domain Guided Dropout layer using six auxiliary datasets for fair comparison.

indicates that the proposed SDL is more effective than DeCov.

Ablation Study Note that SDL differs from DeCov in two aspects: (i) SDL approximates the global covariance by accumulating mini-batch covariance statistics; and (ii) SDL exploits an L_1 instead of L_2 formulation as in DeCov for robustness and correlation sparsity. In order to gain some insight on what contribute to SDL's superior performance, two variants of DeCov (Cogswell et al, 2015), called DeCovGC and DeCovL1, are considered. DeCovGC is DeCov with added accumulating covariance statistic only while DeCovL1 adopts a $L1$ formulation as in SDL. As shown in Table 3.6, both DeCov variants have better results than DeCov (Cogswell et al, 2015) while SDL (with both accumulating covariance statistic and $L1$ loss) achieves the highest performance among them. It suggests that both differences contribute to the effectiveness of SDL.

	CIFAR 10	Market-1501
DeCov	91.62	85.86
DeCovGC	91.86	86.28
DeCovL1	91.90	86.01
SDL	**92.44**	**86.75**

Table 3.6: Ablation study on the advantage of SDL over DeCov. The CIFAR10 classification accuracy (%) and the Market-1501 results are in R1 accuracy (%) under the single query setting.

3.3 Summary

This chapter has presented a novel deep CCA model, term Soft CCA, for the cross-view correlation learning in visual data. Soft CCA provides an efficient and effective solution to deep CCA optimisation by introducing a soft decorrelation loss. Extensive experiments show that the proposed Soft CCA is more effective and scalable than existing deep CCA methods. Compared to exact decorrelation solutions, Soft CCA is easy to implement in contemporary learning frameworks, and therefore is promising for enabling practical use of CCA techniques in the deep learning community. Moreover, we demonstrated that as a by-product, the developed SDL loss can be applied beyond CCA as a general purpose decorrelation loss – to any deep learning task where feature decorrelation is required. As case studies, SDL was shown to outperform alternative decorrelation losses in FAE latent factor disentanglement and CNN object and instance recognition.

Deep CCA has two potential limitations in handling multi-view leaning tasks with data from more than two views. On the one hand, the data of each view requires a specific model for the projection from visual to latent space in CCA. As a result, the model size of CCA will be increased as more views incorporated in a problem. This leads to memory inefficient issue in deep CCA when the number of views is large since an independent deep neural network is deployed for a view. On the other hand, it is harder to train a deep CCA with more views. The learning schedules of different deep networks can be various and coordinating them for optimal global performance is also challenging.

Chapter 4

Deep Factorisation for Multi-View Learning

Deep Canonical Correlation Analysis (CCA) aligns different views via the correlation-based loss and the view-specific deep models. However, deep CCA is inefficient in handling the multi-view data with more than two views since multiple view-specific deep neural networks can end up with a large overall model size. Optimising multiple deep models simultaneously with view-specific inputs is another obstacle in deep CCA training. Moreover, the cross-view alignment only takes place in the embedding space of final layers in deep CCA.

This chapter presents a novel deep factorisation model that is capable of learning the view-invariant latent factors with no additional attribute annotations to handle the multi-view data. Specifically, the proposed deep neural network (DNN) architecture, called Multi-Level Factorisation Net (MLFN), is built on stacking the novel factorisation DNN block, as introduced in Figure 2.3 and the overall architecture is illustrated in Figure 4.1. Different blocks (from bottom to top) contains the latent factors at specific levels (from low to high). By subjecting to the multi-view supervision only, MLFN aims to automatically discover the latent discriminative view-invariant factors at multiple semantic levels and dynamically identify their presence in each visual input. Moreover, a compact feature, called Factor Signature (FS), is generated by aggregating the discriminative information from all levels. It enables an efficient fusion architecture in MLFN to complement the final-layer deep feature. Finally, MLFN is a holistic single DNN architecture based on instance level inputs and labels and avoids the cross-view pairing and view-specified modelling as in deep CCA.

This chapter is structured as follows. Section 4.1 provides the technical details of the pro-

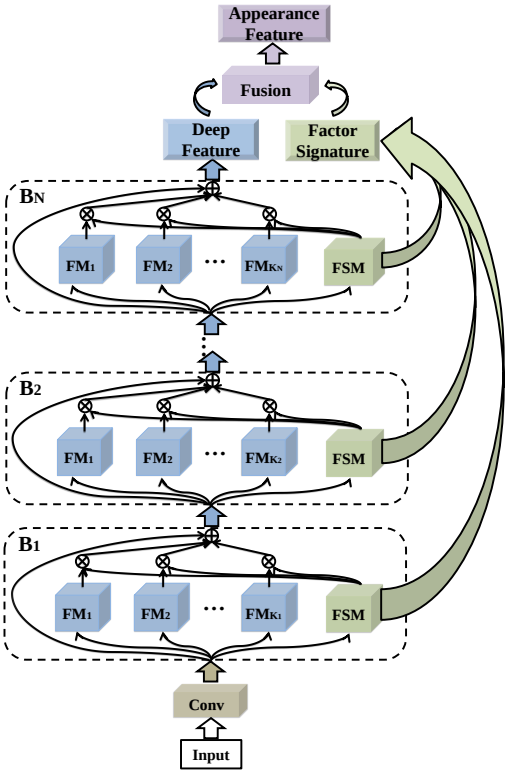


Figure 4.1: Illustration of Multi-Level Factorisation Net (MLFN) Architecture. Best viewed in colour.

posed MLFN. Experimental part is in Section 4.2. Person re-identification (Re-ID) is a typical visual application of cross-view recognition. The effectiveness of the proposed MLFN is mainly demonstrated on a range of Re-ID datasets. Moreover, MLFN also achieves compelling results on the general object categorisation CIFAR-100 dataset (Krizhevsky and Hinton, 2009). Finally, a summary is given in Section 4.3.

4.1 Multi-Level Factorisation Network for Multi-View Latent Factor Discovery

4.1.1 Model Architecture

The proposed MLFN architecture is shown in Figure 4.1, L MLFN blocks are stacked to model L semantic levels. Let B_l denote the lth block, $l \in \{1,...,L\}$ from bottom to top. Within each B_l, there are two key components: multiple Factor Modules (FMs) and a Factor Selection Module (FSM). Each FM is a sub-network with multiple convolutional and pooling layers of its own, powerful enough to model a latent factor at the corresponding level indexed by l. Each block B_l consists of K_l FMs with an identical network architecture. For simplicity, only one input image is considered in the following formulation. Given an image, the output of the ith ($i \in \{1,...,K_l\}$)

FM in B_l is denoted as

$$\boldsymbol{M}_{l,i} \in \mathbb{R}^{H_l \times W_l \times C_l}, \tag{4.1}$$

where $\boldsymbol{M}_{l,i}$ is a feature map with height H_l, width W_l and C_l channels.

Each block B_l also contains a FSM that produces a FM selection vector $\boldsymbol{S}_l \in \mathbb{R}^{1 \times K_l}$. To handle the case where multiple discriminative latent factors are required simultaneously to explain the visual appearance of the input image, within each level, $\boldsymbol{S}_l$ is sigmoid activated,

$$\mathbf{S}_l = \sigma(\bar{\boldsymbol{A}}_l), \tag{4.2}$$

where $l \in \{1,...,L\}$, $\sigma(\cdot)$ is an element-wise sigmoid and $\bar{\boldsymbol{A}}_l$ is the pre-activation output of the FSM.

Thus the factorised representation of an input image at the lth level can be represented as a tuple:

$$\{\boldsymbol{M}_l, \boldsymbol{S}_l\}, \tag{4.3}$$

where $\boldsymbol{M}_l \in \mathbb{R}^{H_l \times W_l \times C_l \times K_l}$ assembles all $\boldsymbol{M}_{l,i}, i \in \{1,...,K_l\}$. The FSM output vector $\boldsymbol{S}_l$ is used for modulating outputs $\boldsymbol{M}_l$ from corresponding FMs. Moreover, shortcut connection is employed by each MLFN block. Therefore, the output of B_l is

$$\tilde{\boldsymbol{Y}}_l = \boldsymbol{M}_l \times_4 \boldsymbol{S}_l + \tilde{\boldsymbol{X}}_l, \tag{4.4}$$

where $\times_4$ denotes the *mode*-4 product of Tensor-matrix multiplication; $\tilde{\boldsymbol{Y}}_l \in \mathbb{R}^{H_l \times W_l \times C_l}$ denotes the output tensor of B_l and $\tilde{\boldsymbol{X}}_l$ is the corresponding input. $\tilde{\boldsymbol{X}}_l$ is from the output of previous block $\tilde{\boldsymbol{Y}}_{l-1}$ and the output of an initial convolutional layer is used as input when $l = 1$.

Factor Signature In order to complement the final-level deep representation $\tilde{\boldsymbol{Y}}_L$ (feature output of B_L) with the factorised representation learned from lower levels, a compact Factor Signature (FS) representation preserving discriminative information from all levels is computed. FS aggregates all FSM output vectors $\boldsymbol{S}_l$, $l \in \{1,...,L\}$. Denoting FS as $\hat{\boldsymbol{S}}$,

$$\hat{\boldsymbol{S}} = [\boldsymbol{S}_1, ..., \boldsymbol{S}_L], \tag{4.5}$$

where $\hat{\boldsymbol{S}} \in \mathbb{R}^{1 \times L}$, $K = \sum_{l=1}^{L} K_l$ represents the feature dimension of $\hat{\boldsymbol{S}}$. The value of L depends on the architecture of MLFN, i.e., both the total number of blocks L and the number of FMs K_l in each block. However, it is independent of the deep feature dimensions in $\tilde{\boldsymbol{Y}}_l$. Therefore, $\hat{\boldsymbol{S}}$ provides a compact multi-level representation even when the deep feature dimension $H_l \times W_l \times C_l$

is large, and when information from all levels is combined. Usually, K is in the order of hundreds and it is much smaller than concatenating all $\tilde{\boldsymbol{Y}}_l$s, which typically results in tens of thousands of dimensions.

Fusion MLFN fuses the deep features $\tilde{\boldsymbol{Y}}_L$ computed from the final block B_L and the Factor Signature (FS) $\hat{\boldsymbol{S}}$. Concretely, $\tilde{\boldsymbol{Y}}_L$ and $\hat{\boldsymbol{S}}$ are first projected to the same feature dimension d with projection function T implemented as a fully connected layer. The final output representation $\boldsymbol{R}$ of MLFN is computed by averaging the two projected features as in Eq. 4.6.

$$\begin{aligned} \boldsymbol{R} &= \frac{1}{2}(\boldsymbol{f}_{\tilde{\boldsymbol{Y}}} + \boldsymbol{f}_{\hat{\boldsymbol{S}}}), \\ &\begin{cases} \boldsymbol{f}_{\tilde{\boldsymbol{Y}}} = T(\tilde{\boldsymbol{Y}}_L, d) \\ \boldsymbol{f}_{\hat{\boldsymbol{S}}} = T(\hat{\boldsymbol{S}}, d) \end{cases} \end{aligned} \tag{4.6}$$

4.1.2 Optimisation

The visual appearance of each input is dynamically factorised into $\{\boldsymbol{M}_l, \boldsymbol{S}_l\}$ at multiple semantic levels[1] in the corresponding MLFN block $B_l, l \in \{1, ..., L\}$, as in Eq. 4.3. Denoting the ith FM in B_l as $F_{l,i}(\cdot)$ and its parameters as $\theta_{l,i}$, then

$$\boldsymbol{M}_{l,i} = F_{l,i}(\tilde{\boldsymbol{X}}_l; \theta_{l,i}). \tag{4.7}$$

The output feature $\tilde{\boldsymbol{Y}}_l$ is computed as in Eq. 4.4. Assuming MLFN is subject to a final loss ℓ and the gradient $\frac{\partial \ell}{\partial \tilde{\boldsymbol{Y}}_l}$ can be acquired. In order to update the parameters $\theta_{l,i}$ in backpropagation, the following gradient is computed,

$$\frac{\partial \ell}{\partial \theta_{l,i}} = \frac{\partial \ell}{\partial \tilde{\boldsymbol{Y}}_l} \frac{\partial \tilde{\boldsymbol{Y}}_l}{\partial F_{l,i}} \frac{\partial F_{l,i}}{\partial \theta_{l,i}}. \tag{4.8}$$

From Eq. 4.4 and Eq. 4.7,

$$\frac{\partial \tilde{\boldsymbol{Y}}_l}{\partial F_{l,i}} = S_{l,i}, \tag{4.9}$$

where $S_{l,i}$ is the FSM output corresponding to the ith FM in B_l. Combining Eq. 4.8 and Eq. 4.9,

$$\frac{\partial \ell}{\partial \theta_{l,i}} = S_{l,i} \frac{\partial \ell}{\partial \tilde{\boldsymbol{Y}}_l} \frac{\partial F_{l,i}}{\partial \theta_{l,i}}, \tag{4.10}$$

where $\frac{\partial \ell}{\partial \tilde{\boldsymbol{Y}}_l}$ is back propagated from higher levels and $\frac{\partial F_{l,i}}{\partial \theta_{l,i}}$ is the gradient of an FM w.r.t its parameters. $S_{l,i}$ comes from the corresponding FSM. It dynamically indicates the contribution of $F_{n,i}$ in processing an input image.

[1] Figure 4.5 illustrates the learned multiple levels visual semantic factors by the proposed MLFN. Relevant explanation can be found at **What is Learned** paragraph in Section 4.2.4

$S_{l,i}$ will be close to 1 if the latent factor represented by $\boldsymbol{M}_{l,i}$ is identified to be present in the input. In this case, the impact of this input is fully applied on $\theta_{l,i}$ to adapt the corresponding FM. On the contrary, when $S_{l,i}$ is close to 0, it means the input only holds irrelevant or opposite latent factors to $\boldsymbol{M}_{l,i}$. Therefore, the parameters in the corresponding FM are unchanged when training with this input as $S_{l,i} \approx 0$ stops the update.

The factor selection vectors $\boldsymbol{S}_l$ (Eq. 4.2) play a key role in MLFN during both training (as analysed above) and inference (providing the factor signature). Learning discriminative FSMs would be hard if trained with gradients back propagated through many blocks from the top. This is because the supervision from the loss would be indirect and weak for the FSMs at the bottom levels. However, because our final feature output $\boldsymbol{R}$ is computed by fusing the final-block output $\tilde{\boldsymbol{Y}}_L$ with the FS $\hat{\boldsymbol{S}}$ (Eq. 4.6), and the FS is generated by concatenating all FSM output vectors, the supervision flows from the loss down to every FSM via direct shortcut connections (Figure 4.1). Thus our FSMs are deeply supervised (Lee et al, 2015; Jin et al, 2016) to ensure that they are discriminative, but without the increase in parameters that would be required for deep supervision of conventional deep features.

MLFN for Person Re-ID The training procedure of MLFN for Person Re-ID follows the standard identity classification paradigm (Xiao et al, 2016; Zhao et al, 2017a; Sun et al, 2017) where each person's identity is treated as a distinct class for recognition. A final fully connected layer is added above the representation $\boldsymbol{R}$ that projects it to a dimension matching the number of training classes (identities), and the cross-entropy loss is used. MLFN is then end-to-end trained. It discovers latent factors with no supervision other than person identity labels for the final classification loss. During testing, appearance representations $\boldsymbol{R}$ (Eq. 4.6) are extracted from gallery and probe images, and the L2 distance is used for matching.

4.2 Results and Analysis

4.2.1 Datasets and Settings

Datasets Person Re-identification (Re-ID) aims to match people across multiple surveillance cameras with non-overlapping views. As shown in Figure 4.2, there are usually more than six camera views in the large-scale person Re-ID datasets. The visual appearance of a person across different views can change drastically due to the distinctive view characteristics such as illumination, background, camera angle and human pose. Three person Re-ID benchmarks, Market-

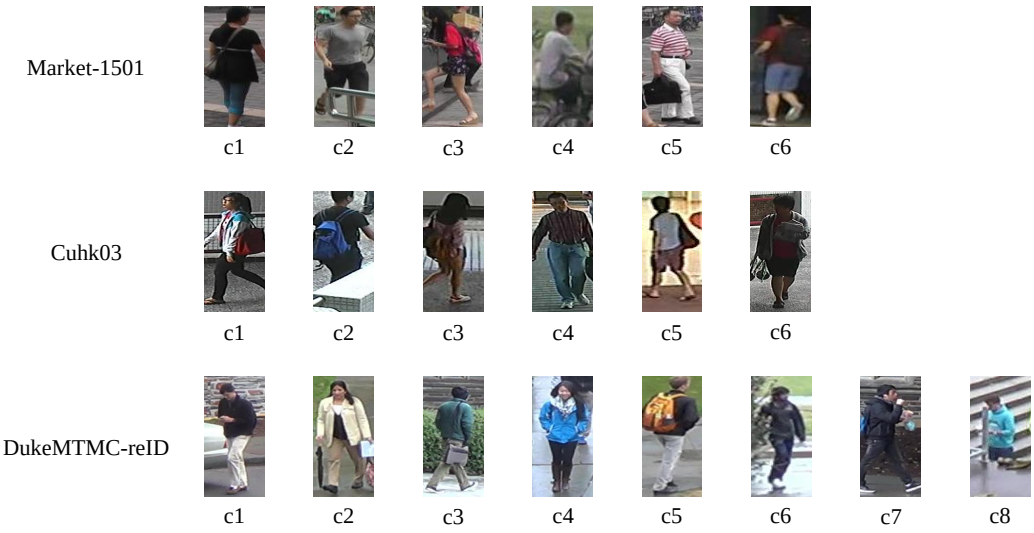


Figure 4.2: Three large scale person Re-ID datasets, Market-1501, CUHK03 and DukeMTMC-reID, are shown with samples from different camera views.

1501 (Zheng et al, 2015), CUHK03 (Li et al, 2014) and DukeMTMC-reID (Zheng et al, 2017) are used for evaluation. **Market-1501** (Zheng et al, 2015) has 12,936 training and 19,732 testing images with 1,501 identities in total from 6 cameras. Deformable Part Model (DPM) (Felzenszwalb et al, 2010) is used as the person detector. The standard training and evaluation protocols in (Zheng et al, 2015) where 751 identities are used for training and the remaining 750 for testing is followed. **CUHK03** (Li et al, 2014) consists of 13,164 images of 1,467 people. Both manually labelled and DPM detected person bounding boxes are provided. Two experimental settings are adopted on this dataset. The first setting, denoted as CUHK03 Setting 1, is the 20 random train/test splits used in (Li et al, 2014) which selects 100 identities for testing and training with the rest. Results on the more challenging yet more realistic detected person bounding boxes are reported under this setting. The other setting, denoted as CUHK03 Setting 2, was proposed in (Zhong et al, 2017). It is more challenging than Setting 1 with less training data. In particular, 767 identities are used for training and the remaining 700 identities for testing. **DukeMTMC-reID** (Zheng et al, 2017) is the Person Re-ID subset of the DukeMTMC Dataset (Ristani et al, 2016). There are 16,522 training images of 702 identities, 2,228 query images and 17,661 gallery images of the other 702 identities. Manually labelled pedestrian bounding boxes are provided. Our experimental protocol follows that of (Zheng et al, 2017).

In addition to the Re-ID datasets, an object category classification dataset, **CIFAR-100** (Krizhevsky and Hinton, 2009), is used to show that our MLFN can also be applied to other general recognition problems. CIFAR-100 (Krizhevsky and Hinton, 2009) has 60K images with 100 classes

MLFN Block	#FSM Layer Ouputs
1-3	128, 64, 32
4-7	256, 128, 32
8-13	512, 128, 32
14-16	512,128, 32

Table 4.1: Architecture details of FSM modules in different MLFN blocks. 16 indicates the last block of MLFN.

with 600 images in each class. 50K images are used for training and the remaining for testing.

Evaluation Metrics The Cumulated Matching Characteristics (CMC) curve is used to evaluate the performance of Re-ID methods. Due to space limitation and for easier comparison with published results, we only report the cumulated matching accuracy at selected ranks in tables rather than plotting the actual curves. Note that mean average precision (mAP) is also used as suggested in (Zheng et al, 2015) to evaluate the performance. For CIFAR100, the error rate is used.

MLFN Architecture Details For Person Re-ID tasks, sixteen blocks ($L = 16$) are stacked in MLFN. Within each building block, 32 FMs are aggregated as in (Xie et al, 2016b). Correspondingly, a 32-D FSM output vector is generated within each MLFN block. As a result, the FS dimension $K = 512$ (32 FMs $\times$16 blocks). The final feature dimension of $\boldsymbol{R}$, d is set to 1024. For the object categorisation task CIFAR-100 (Krizhevsky and Hinton, 2009), the MLFN depth is reduced in order to fit the memory limitation of a single GPU. The number of blocks is reduced to 9 which results in $K = 288$.

The proposed MLFN architecture consists of 16 MLFN Blocks/Layers. A Factor Selection Module (FSM) is included in each Block. The FSM networks used in this paper are all three-layered Multiple Layer Perceptron (MLP). Global Average Pooling (GAP) is applied on the input of FSM. Batch Normalisation and Relu are used to activate each layer's output. Architecture details are shown in Table 4.1.

Data Augmentation The input image size is fixed to 256×128 for all person Re-ID experiments. Left-right flip augmentation is used during training. For CIFAR-100, training images are augmented as in (He et al, 2016a). No data augmentation is used for testing.

Optimisation Settings All person Re-ID models are fine-tuned on ImageNet (Deng et al, 2009) pre-trained networks. The Adam (Kingma and Ba, 2014) optimiser is used with a mini-

	SQ		MQ	
	R1	mAP	R1	mAP
Spindle (Zhao et al, 2017a)	76.9	-	-	-
HP-net (Liu et al, 2017)	76.9	-	-	-
OIM (Xiao et al, 2017)	82.1	-	-	-
Re-rank (Zhong et al, 2017)	77.1	63.6	-	-
DPA (Zhao et al, 2017c)	81.0	63.4	-	-
SVDNet (Sun et al, 2017)	82.3	62.1	-	-
DaF (Yu et al, 2017b)	82.3	72.4	-	-
ACRN (Schumann and Stiefelhagen, 2017)	83.6	62.6	-	-
Context (Li et al, 2017a)	80.3	57.5	86.8	66.7
JLML (Li et al, 2017c)	83.9	64.4	89.7	74.5
LSRO (Zheng et al, 2017)	84.0	66.1	88.4	76.1
SSM (Bai et al, 2017)	82.2	68.8	88.2	76.2
DML (Zhang et al, 2018)	87.7	68.8	91.7	77.1
DPFL (Yanbei et al, 2017)	88.6	72.6	92.2	80.4
MLFN	**90.0**	**74.3**	**92.3**	**82.4**

Table 4.2: Results (%) on Market-1501. −: not reported.

batch size of 64. Initial learning rate is set to 0.00035 for all Re-ID datasets except CUHK03 setting 2 (Zhong et al, 2017) with 0.0005. Similarly, Training iterations are 100k for all Re-ID datasets except CUHK03 setting 2 (Zhong et al, 2017) for which it is 75k. For CIFAR, the initial learning rate is set to 0.1 with a decay factor 0.1 at every 100 epochs and Nesterov momentum of 0.9. SGD optimisation is used with a 256 mini-batch size on a K80 GPU for 307 epochs training.

4.2.2 Person Re-ID Results

Results on Market-1501 Comparisons between MLFN and 14 state-of-the-art methods on Market-1501 (Zheng et al, 2015) are shown in Table 4.2. SQ and MQ correspond to the single and multiple query setting respectively. The results show that our MLFN achieves the best performance on all evaluation criteria under both settings. It is noted that: (1) The gaps between our results and those of the two models (Zhao et al, 2017a; Liu et al, 2017) that attempt to fuse multi-level features are significant: 13.1% R1 accuracy improvement under SQ. This suggests that our fusion architecture with deep supervision is more effective than the handcrafted architectures

	R1	mAP
LSRO (Zheng et al, 2017)	67.7	47.1
OIM (Xiao et al, 2017)	68.1	-
APR* (Lin et al, 2019)	70.7	51.9
ACRN (Schumann and Stiefelhagen, 2017)	72.6	52.0
SVDNet (Sun et al, 2017)	76.7	56.8
DPFL (Yanbei et al, 2017)	79.2	60.6
MLFN	**81.0**	**62.8**

Table 4.3: Results (%) on DukeMTMC-reID. *: Arxiv paper.

with manual layer selection in (Zhao et al, 2017a; Liu et al, 2017), which require extra effort but may lead to suboptimal solutions. (2) The best model that uses attribute annotation (Schumann and Stiefelhagen, 2017) also yields inferior results (SQ 83.6 vs 90.0 for R1 and 62.6 vs 74.3 for mAP), despite the fact that more supervision was used. This indicates that the automatically discovered latent factors at multiple levels in MLFN provides a more discriminative representation. (3) The closest competitor, DPFL uses multiple network branches to model image input scaled to different resolutions, which is orthogonal to our approach and can be easily combined to improve our performance further.

Results on DukeMTMC-reID Person Re-ID results on DukeMTMC-reID (Zheng et al, 2017) are given in Table 4.3. This dataset is challenging because the person bounding box size varies drastically across different camera views, which naturally suits the multi-scale Re-ID models such as DPFL (Yanbei et al, 2017). The results show that MLFN is 1.8% and 2.2% higher than the prior state-of-the-art DPFL (Yanbei et al, 2017) on R1 and mAP metrics respectively. This indicates that even without explicitly extracting features from input images scaled to different resolutions, by fusing features from multiple levels (blocks in MLFN), it can cope with large scale changes to some extent.

Results on CUHK03 Table 4.4 shows results on CUHK03 Setting 1 when detected person bounding boxes are used for both training and testing. MLFN achieves the best result, 82.8%, under this setting. Note that DGD (Xiao et al, 2016), Spindle Net (Zhao et al, 2017a) and HP-net (Zhao et al, 2017a) were trained with the JSTL setting (Xiao et al, 2016) where additional data in the form of six Re-ID datasets were used. They also used mixed labelled and detected

	R1
DGD$^\sharp$ (Xiao et al, 2016)	75.3*
Spindle$^\sharp$ (Zhao et al, 2017a)	88.5*
HP-net$^\sharp$ (Liu et al, 2017)	91.8*
LSRO† (Zheng et al, 2017)	84.6
OIM (Xiao et al, 2017)	77.5
JLML (Li et al, 2017c)	80.6
SVDNet (Sun et al, 2017)	81.8
DPFL (Yanbei et al, 2017)	82.0
MLFN	**82.8**/89.2*

Table 4.4: Results (%) on CUHK03 Setting 1 (Li et al, 2014). $^\sharp$ indicates using external Re-ID data (JSTL setting (Xiao et al, 2016)). Results with * are obtained with the same setting in (Xiao et al, 2016). † indicates GAN images generated from the Market-1501 dataset are used.

bounding boxes for both training and testing. Following the multi-bounding box setting, even without using auxiliary training data as in JSTL, the accuracy of MLFN jumps from 82.8% to 89.2%. Similarly, LSRO (Zheng et al, 2017) used external Re-ID datasets for training, thus gaining an advantage.

The results in Table 4.5 correspond to CUHK03 Setting 2, which is a harder and newer setting with less reported results. Clear gaps are now shown between MLFN and DPFL (Yanbei et al, 2017): The rank 1 (R1) performance of MLFN is more than 11% higher using either labelled or detected person images. This result suggests that the advantage of MLFN is more pronounced given less training data. Similar performance jumps are also observed using the mAP metric.

	Labelled		Detected	
	R1	mAP	R1	mAP
DaF (Yu et al, 2017b)	27.5	31.5	26.4	30.0
Re-rank (Zhong et al, 2017)	38.1	40.3	34.7	37.4
SVDNet (Sun et al, 2017)	40.9	37.8	41.5	37.3
DPFL (Yanbei et al, 2017)	43.0	40.5	40.7	37.0
MLFN	**54.7**	**49.2**	**52.8**	**47.8**

Table 4.5: Results (%) on CUHK03 Setting 2.

4.2.3 Object Categorisation Results

We next evaluate whether our MLFN is applicable to more general object categorisation tasks by experimenting on CIFAR-100. The results are shown in Table 4.6. For direct comparison we reproduce results with ResNet (He et al, 2016b) and ResNeXt (Xie et al, 2016b) of similar depth and model size to our MLFN. The improved result over ResNeXt shows that our dynamic factor module selection and factor signature feature bring clear benefit. MLFN also beats DualNet (Hou et al, 2017), another representative recent ResNet-based model that fuses two complementary ResNet branches as in an ensemble, thus doubling in model size. Note that for distinguishing different object categories, e.g., dog and bird, low-level factors such as colour and texture are often less useful as for instance classification problems such as person Re-ID. However, this result suggests that discriminative latent factors still exist in multiple levels for object categorisation and can be discovered and exploited by our MLFN.

	Error Rates (%)
DualNet (Hou et al, 2017)	27.57
ResNet (He et al, 2016b)	30.21
ResNeXt (Xie et al, 2016b)	29.03
MLFN	**27.21**

Table 4.6: Results on CIFAR-100 datasets.

4.2.4 Further Analysis

Ablation Study Recall that our MLFN discovers multiple discriminative latent factors at each semantic level, by aggregating FMs with identical structures within each block B_l. The FSM output vectors $\boldsymbol{S}_l$ enable dynamic factorisation of an input image into distinctive latent attributes, and these are aggregated over all blocks into a compact FS feature ($\hat{\boldsymbol{S}}$) for fusion (Eq. 4.6) with the conventional (final-block) deep feature $\tilde{\boldsymbol{Y}}_L$ to produce the final representation $\boldsymbol{R}$. To validate the contributions of each component, we compare: **MLFN:** Full model. **MLFN-Fusion:** MLFN using dynamic factor selection, but without fusion of the FS feature. **ResNeXt:** When the FSMs are removed so all FMs are always active, our model becomes ResNeXt (Xie et al, 2016b). **ResNet:** When the sub-networks at each level of ResNeXt are replaced with one larger holistic residual module, ResNet (He et al, 2016b) is obtain. A comparison of these models on all three

Datasets	Market-1501				CUHK03				DukeMTMC-reID	
Methods	SQ		MQ		Labelled		Detected		R1	mAP
	R1	mAP	R1	mAP	R1	mAP	R1	mAP		
ResNet	84.3	66.0	89.6	76.1	41.7	37.9	43.5	38.6	71.6	48.6
ResNeXt	88.0	69.8	91.3	79.0	43.8	38.7	43.1	38.0	75.7	54.1
MLFN-Fusion	87.9	70.8	91.7	80.2	47.1	42.5	47.1	41.0	78.7	58.4
MLFN	**90.0**	**74.3**	**92.3**	**82.4**	**54.7**	**49.2**	**52.8**	**47.8**	**81.0**	**62.8**

Table 4.7: Ablation Results on three Person Re-ID datasets. CUHK03 results were obtained under Setting 2.

person Re-ID datasets is shown in Table 4.7. It shows that MLFN is consistently better than the stripped-down versions on all datasets, and each new component contributed to the final performance: The margin between MLFN and MLFN-Fusion shows the importance of including the latent factor descriptor FS in the representation and suggests that the FS feature is complement to the final-block feature $\tilde{\boldsymbol{Y}}_L$, and the margin between MLFN$-$Fusion and ResNeXt shows the benefit of dynamic module selection.

MLFN Architecture Parameter Selection The number of blocks (L) in MLFN is set to 16 follows the ResNeXt-50 (Xie et al, 2016b). The FS dimension K depends on L and the number of FMs at each MLFN block. These are set, without tuning, so that the model is of a comparable size to ResNeXt-50 (Xie et al, 2016b) for direct comparison. On GTX1080 GPU, the runtime is similar: MLFN (0.81s/batch) and ResNeXt (0.78s/batch), and so is the GPU memory consumption. The final feature dimension d of MLFN is 1024 since it is the widely used feature dimension for Person Re-ID (Sun et al, 2017). The impacts of ds on the re-id performance are illustrated as in Figure 4.3. It can be seen that the performance is consistently good when $d > 512$.

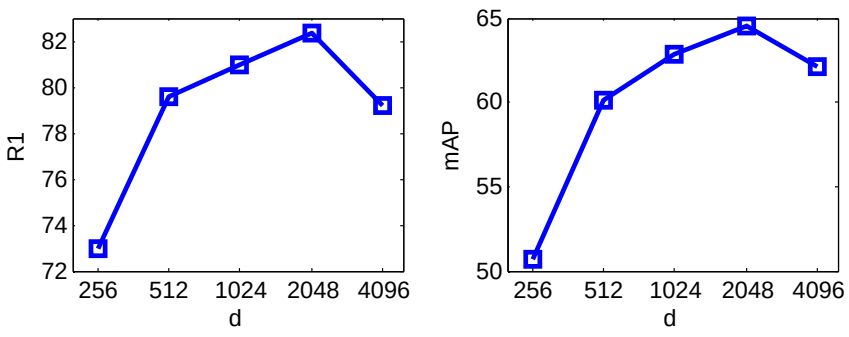


Figure 4.3: Sensitivity to dimemsion d. DukeMTMC-reID (Zheng et al, 2017) is used.

Efficacy of Re-ID with Factor Signature Alone For solely FS-based matching, a binary SVM is trained based on the absolute difference of paired FS to predict whether they belong to the same person or not. SVM scores of testing pairs are then computed for recognition. The corresponding results on Market-1501 are reported in Table 4.8. It shows that, compared with the results in Table 4.2, the result of FS only is already comparable with the state-of-the-art.

	SQ		MQ	
	R1	mAP	R1	mAP
$\hat{\boldsymbol{S}}$	81.0	58.9	88.0	68.8

Table 4.8: Market-1501 (Zheng et al, 2015) Re-ID performance (%) with $\hat{\boldsymbol{S}}$ (FS) only.

Discovered Latent Factors are Predictive of Attributes What do the discovered latent factors represent? We hypothesise that despite not being trained with any manually annotated attributes, FS ($\hat{\boldsymbol{S}}$) is identifying latent data-driven attributes present in the data; these latent attribute may overlap or correlate with human-defined semantic attributes. To validate this, SVMs are then trained based on $\hat{\boldsymbol{S}}$ only to predict ground-truth manually annotated attributes in Market-1051 and DukeMTMC-reID. Results based on the final representation $\boldsymbol{R}$ from MLFN are also reported. Finally, these are compared to APR (Lin et al, 2019), which is end-to-end trained based on attribute supervision.

On Market-1051, MLFN-$\hat{\boldsymbol{S}}$ and APR (Lin et al, 2019) obtain the same performance of 85.33%. MLFN-$\boldsymbol{R}$ further improves to 87.50%. On DukeMTMC-reID, 82.30% and 83.58% are achieved by MLFN-$\hat{\boldsymbol{S}}$ and MLFN-$\boldsymbol{R}$ respectively, which are better than APR's 80.12%. These results thus show that our low-dimensional MLFN-$\hat{\boldsymbol{S}}$ alone can be more effective in attribution prediction than APR. Remind that MLFN is trained without annotated attributes while APR network is designed for supervised attribute learning. This shows that our architecture is well suited for extracting semantic attribute related information automatically.

The relations between FS and Attributes can be further illustrated by Figure 4.4 where the predicted attributes using our FS feature and the human labelled attributes are compared. For each person image, 35 binary attributes are annotated by human annotators on the identity level, that is, different images of the same person would have the identical attribute vectors regardless whether those attributes are visually observable in the images. These attributes form different groups and within each group, they are mutually exclusive. For example, female and male form

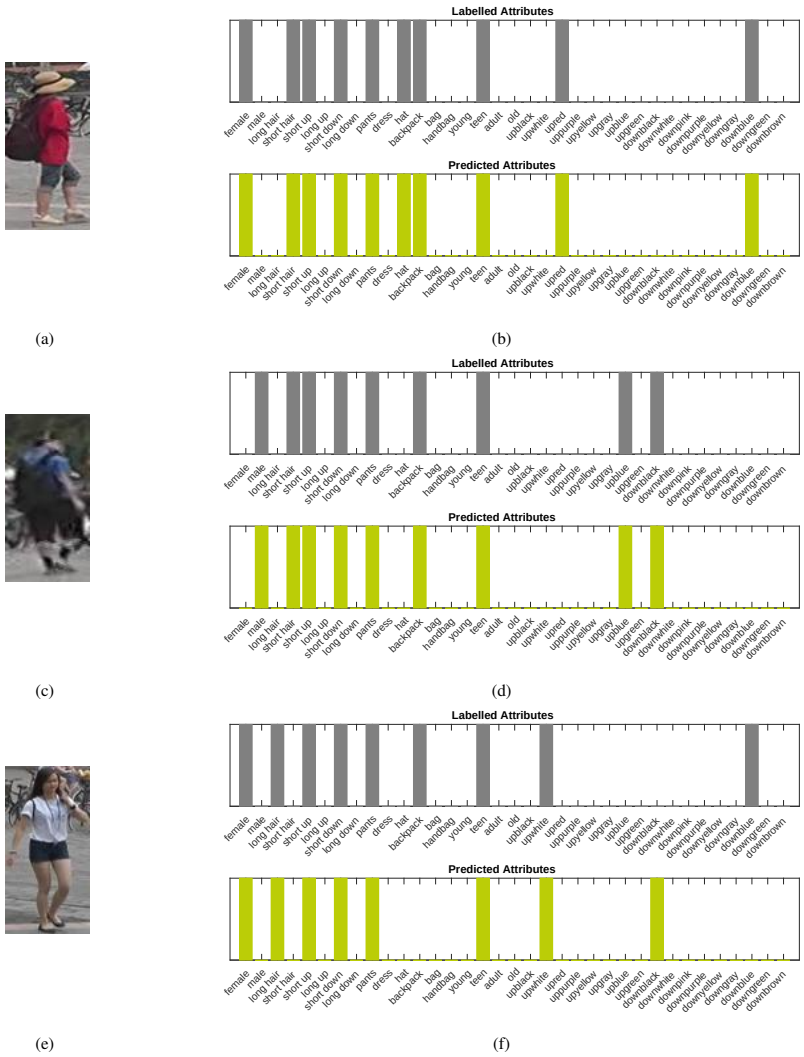


Figure 4.4: Examples of attribute prediction using our factor signature (FS) feature. Market-1501 dataset is used. Best viewed in colour.

one group, and young, teen, adult, old form another. Some attributes are thus subjective, e.g., no ground-truth age is known and there is no clear definition of what 'young' entails.

Figure 4.4(a) shows an example where our FS feature can be used to correctly predict all the attributes with SVM classifiers. In this example, although the big hat occludes the face and part of the hair of the person, the colour of the top and the shoe style give away the fact that this a female. A harder example is shown in Figure 4.4(c). This time the image is a bit blurred and the viewpoint is from the back. However, our FS feature can still predict all the attributes correctly. Our FS feature based prediction makes two mistakes for the person image shown in Figure 4.4(e). Specifically, the backpack attribute is missed and the lower-body garment colour

Figure 4.5: Four groups of images corresponding to highest (first row) and lowest (second row) values of four FSM outputs $S_{l,i}$, from bottom to top level respectively. Best viewed in colour.

is predicted to be black rather than blue. Both mistakes are understandable. For the backpack, since the frontal view is shown and the backpack has very thin straps, this attribute can be easily missed even by human (the human annotator labelled this because s/he had access to multiple views of this person including a back view where the backpack is clearly visible). As for the blue vs black for the lower-body cloth, it seems to be a close call even for humans.

What is Learned To visualise the latent discriminative appearance factors learned by MLFN, each element of the FSM output vector is ranked, denoted as $S_{l,i}$, with all testing samples in Market-1501 (Zheng et al, 2015) as inputs. Person images with the highest and lowest twenty values of each $S_{l,i}$ are recorded. Figure 4.5 shows four example sets of such images from different element $i, i \in \{1,..,K_L\}$ and blocks $l, l \in \{1,...,L\}$. Clear visual semantics can be seen from both the highest and lowest FSM output value image clusters in each group. And as expected, as the block index number l increases, the semantic level of the latent factors captured at the corresponding blocks gets higher, i.e., they evolve from colour and texture related factors to clothes style and gender related ones. This is achieved despite that *no attribute supervision is used in training MLFN*. It is also interesting to note that visual characteristics conveyed by images with the highest FSM output values are complementary or opposite to those of lowest ones from the same group. For example, highest value images in $S_{2,29}$ contain green colour,

while lowest value images contain the complementary colour red. High value in $S_{7,31}$ encodes cold colours while low value encodes warm colours. Highest values in $S_{10,23}$ reflect textures while lowest ones mean large untextured colour blocks are detected. Images of men select with high confidence $S_{15,29}$, while images of females depress its value.

4.3 Summary

This chapter presents a deep factorisation model for learning with the supervised multi-view, i.e., more than two views, data. Specifically, a novel deep neural network (DNN) architecture, Multi-Level Factorisation Net (MLFN), has been proposed. It learns to discover and dynamically identify discriminative latent factors in multi-view visual inputs. The factors computed at different levels of the network correspond to hidden attributes of different semantic levels. When the selection of the factors are used as a feature and fused with the conventional deep feature, a powerful view-invariant visual appearance feature representation is obtained. MLFN is applied to the person Re-ID, a challenging cross-view recognition problem. MLFN consistently outperforms the state-of-the-art deep models on three largest Re-ID datasets and shows promising performance on a more general object categorisation task.

The cross-view recognition algorithms proposed in Chapter 3 and this chapter requires the fully supervised multi-view data for model training. However, labelling the multi-view visual data can be time-consuming and tedious. Multi-view data can lack annotations on specific views. And the supervised learning methods cannot exploit the data from the unlabelled views. The next chapter deals with a challenging multi-view data setting with an unlabelled dataset as the target view/domain. Another labelled dataset with the relevant task is exploited as the source domain. Each domain corresponds to a multi-view dataset, and the label spaces across domains are disjoint. The primary objective is to improve the performance of the target domain with data from both source and target datasets for the model training.

Chapter 5

Unsupervised Domain Adaptive Multi-View Learning

The previous two chapters propose the supervised multi-view learning algorithms, Soft CCA (Chapter 3) and MLFN (Chapter 4), for different types of fully labelled multi-view visual data. However, the supervision of multi-view data are expensive to acquire. In many application scenarios, the data samples of the views are unlabelled. In this chapter, a challenging unsupervised multi-view learning setting, called unsupervised domain adaptive (UDA) multi-view learning, is investigated. Specifically, an unlabelled multi-view dataset is considered as the target view/domain. To enhance the performance of the target domain, the source domain, a fully supervised multi-view dataset with relevant tasks, is utilised. However, training the supervised learning algorithms on the source domain and directly applied it to the target one usually results in unsatisfied performance due to the existence of domain gaps. Most existing approaches for unsupervised domain adaptation are domain alignment methods. They are based on the assumption of a shared label space across different domains. Therefore, different domains should be aligned in either the raw image space (Wei et al, 2018) or a feature embedding (Ganin and Lempitsky, 2015). Nevertheless, the UDA multi-view learning setting holds an opposite assumption: the label spaces across domains are disjoint. The domain alignment methods are thus not suitable under this setting.

In this chapter, we propose to tackle the UDA multi-view setting by *multi-task learning* a shared deep feature embedding space. Different domains make distinctive contributions to the space learning. The learning task of source domain is supervised label prediction while the target objective can be specified under different unsupervised learning assumptions. Therefore, the

proposed multi-task framework is totally different from existing methods in that no attempt for domain alignment is made. Specifically, two different assumptions on target domain are made and thus result in distinctive models for UDA multi-view learning problems. The first model, called Common Factorised Space Model (CFSM), assumes that recognition should be performed in a shared latent factor space for both domains where each factor can be interpreted as latent attribute. An unsupervised factorisation loss is proposed to serve this purpose on target domain. The second model is based on a simple and sound assumption, the target multi-view data instances form clusters in the shared space and each cluster potentially corresponds to an unknown entity/object. To fulfill this objective, a novel deep clustering method, Stochastic Inference for Deep Clustering (SIDC), is developed.

This chapter is organised as below. The CFSM is presented first in Section 5.1. As show in Section 5.2 that CFSM is widely applicable to many transfer learning tasks besides UDA multi-view learning. The details of SIDC are then described in Section 5.3. Its performance on an UDA multi-view learning problem, i.e., the UDA person Re-ID, are reported in Section 5.4. Finally, a summary of the proposed methods is given in Section 5.5.

5.1 Common Factorised Space Model

In this section, the Common Factorised Space Model (CFSM) is the main focus. It provides a simple yet effective multi-task training solution for UDA multi-view learning problems. To discover a shared latent factor space across domains for recognition, an unsupervised factorisation loss is proposed. Moreover, a novel graph Laplacian-based loss is derived to better exploit the more aligned and discriminative supervision from higher-level to improve deep feature learning. Besides the UDA multi-view learning problems, the CFSM is effective on other tasks such as the conventional UDA.

5.1.1 Methodology

Definition and Notation For UDA multi-view learning, there is a source (labelled) domain S and a target (unlabelled or partially labelled) domain T. The key characteristic of UDA multi-view learning is the disjoint label space, i.e., the source Y_S and target Y_T label spaces are potentially disjoint: $Y_S \cap Y_T = \varnothing$. Instances from source/target domains are denoted X_S and X_T respectively. The combined inputs $\{X_S, X_T\}$ are denoted as X.

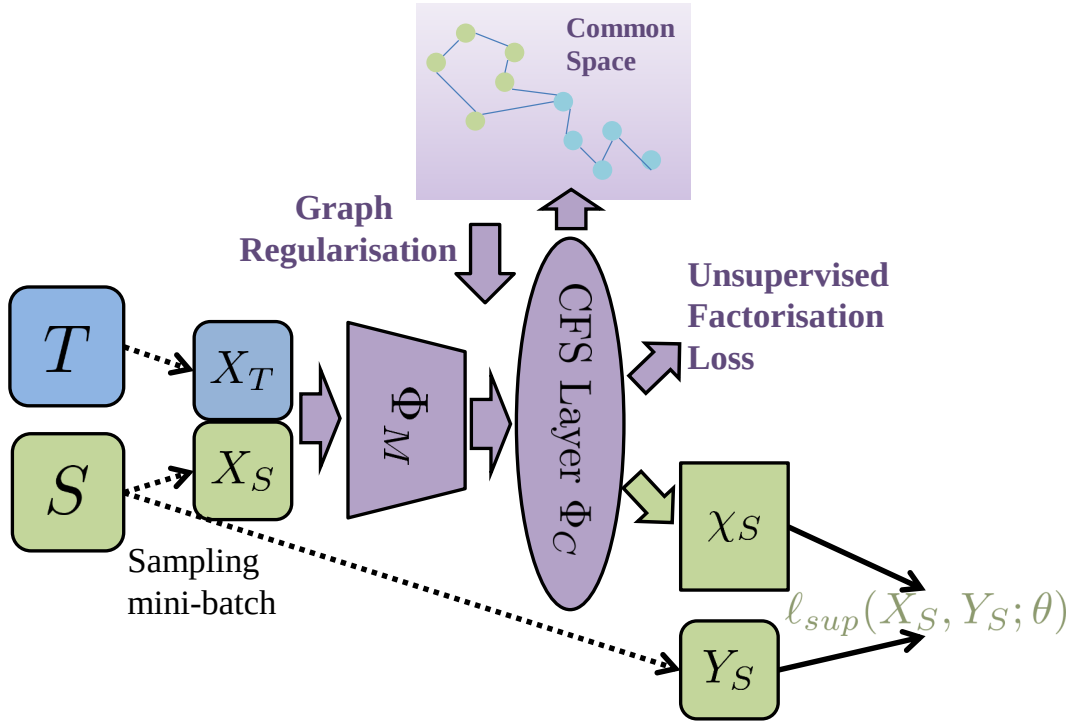


Figure 5.1: Different colours corresponding to different data streams. Green indicates source data. Blue is used for target data. Purple means joint data from both source and target domains. The parameters of CFSM are θ_M, θ_C and θ_S corresponding to the feature extractor Φ_M, CFS layer Φ_C and source classifier χ_S. θ_S is learned with source data using supervised loss while θ_M, θ_C are estimated using data from both domains with all losses and regularisations.

Model Architecture The proposed model architecture consists of three modules, a feature extractor $F = \Phi_M(X)$ that can be any deep neural network and is shared between all domains. This is followed by a fully connected layer and sigmoid activation σ, which define the Common Factorised Space (CFS) layer. This provides a representation of dimension d_C, $\boldsymbol{f}_C = \Phi_C(\cdot) = \sigma(\boldsymbol{W}\Phi_M(\cdot) + \boldsymbol{b})$. Recall that the goal of CFS is to learn a latent factor (low-entropy) representation for both source and target domains. The sigmoid activation means that the layer's scale is $\boldsymbol{f}_C \in (0,1)^{d_C}$, so activations near 0 or 1 can be interpreted as the corresponding latent factor being present or absent. To encourage a near-binary representation, unsupervised factorisation loss is applied. For the labelled source domain only, the pre-activated $\boldsymbol{f}_C$ are then classified by softmax classifier χ_S with cross-entropy loss. The overall architecture is illustrated in Figure 5.1.

Regularised Model Optimisation The parameters of the proposed CFSM are $\theta := \{\theta_M, \theta_C, \theta_S\}$ including parameters of the feature extractor Φ_M, CFS layer Φ_C and source classifier χ_S. Both the labelled source $\{X_S, Y_S\}$ and unlabelled target data X_T are used in the multi-task model training procedure.

Firstly, the labelled source data $\{X_S, Y_S\}$ contributes to the model training as a supervised learning task with a loss $\ell_{sup}(X_S, Y_S; \theta)$ which is a conventional cross-entropy. However, such loss is inapplicable to the target domain data since no supervision is provided.

To adapt the knowledge across domains, the unlabelled target data plays the key role in model training. Therefore, unsupervised domain adaptation losses/regularisations are required to enable the multi-task learning with target domain. As mentioned before, conventional UDA losses/regularisations are not applicable in the UDA multi-view learning problem since the label space across domains are disjoint. Therefore, we proposed the Common Factorised Space (CFS) by using the CFS layer Φ_C. A low-entropy loss is used to regularise its model learning (parameter θ_C). On the other hand, the UDA multi-view learning problem usually relies on the feature representation for retrieval tasks, e.g., in person Re-ID. More regularisations are thus required on the learning of feature extractor Φ_M (with parameter θ_M). As illustrated in Figure 5.1, two unsupervised regularisations are exploited in our CFSM based on the data samples from both domains.

Low-Entropy Regularisation: Unsupervised Adaptation Firstly, the definition of the low-entropy regulariser on the CFS is discussed. The sigmoid activated outputs $\boldsymbol{f}_C$ from CFS layer Φ_C can be interpreted as multi-label predictions on latent factors. The uncertainty measure for label prediction can be defined by using its entropy,

$$\begin{aligned} &-\sum_{i=1}^{N} < \boldsymbol{f}_{C,i}, \log(\boldsymbol{f}_{C,i}) > \\ =&-\sum_{i=1}^{N} < \Phi_C(\mathbf{x}_i), \log(\Phi_C(\mathbf{x}_i)) > \end{aligned} \tag{5.1}$$

where $\boldsymbol{f}_{C,i}$ denotes the common factor representation $\Phi_C(\boldsymbol{x}_i)$ of instance $\boldsymbol{x}_i \in X$. This is applied on both source and target data, so N is the number of instances in both datasets. $\log(\cdot)$ is applied element-wise, and $< \cdot, \cdot >$ is vector inner product. According to the low-uncertainty criterion (Carlucci et al, 2017), optimising such prior can be achieved by minimising this uncertainty measure. Eq. 5.1 is thus the regulariser corresponding to the low-entropy prior. Specifically, this loss biases the representation F_C to contain more certain predictions, e.g., closer to 0 or 1 for each discovered latent factor. Therefore, it is denoted as unsupervised factorisation loss.

In summary, the low-entropy regulariser on CFS is built upon the assumption that the two domains share a set of latent attributes and that if a source classifier is well adapted to the target, then the presence/absence of these attributes should be certain for each instance. Therefore, it

essentially generalises the low-uncertainty principle (widely used in existing unsupervised and semi-supervised learning literature) to the disjoint label space setting.

Graph Regularisation: Robust Feature Learning The second prior is regularising the feature extractor Φ_M. The unique property of our setup so far is that the knowledge transfer into the target domain is via the CFS layer; therefore we are interested in ensuring that the feature extractor network extracts features whose similarity structure reflects that of the latent factors in the CFS layer. Unlike conventional graph Laplacian losses that regularise higher-level features with a graph built on lower-level features (Belkin et al, 2006; Zhu, 2005), we do the reverse and regularise the feature extractor Φ_M to reflect the similarity structure in $\boldsymbol{f}_C$. This is particularly important for applications where the target problem is retrieval, because deep features $\boldsymbol{f} = \Phi_M(\cdot)$ are used as an image representation.

The proposed graph loss is expressed as

$$\mathrm{Tr}(\boldsymbol{f}^T \Delta_{\boldsymbol{f}_C} \boldsymbol{f}), \tag{5.2}$$

where $\Delta_{\boldsymbol{f}_C}$ is the graph Laplacian (Cai et al, 2010b) built on the common space features $\boldsymbol{f}_C$.

Summary We unify the proposed model architecture $\theta := \{\theta_M, \theta_C, \theta_S\}$ with source $\{X_S, Y_S\}$ and target $\{X_T\}$ data for UDA multi-view problems under the multi-task learning framework. This decomposes into a standard supervised term (with source data only) and data-driven priors for the CFS layer and feature extraction module. They correspond to supervised loss $\ell_{sup}(X_S, Y_S; \theta)$, unsupervised factorisation loss (Eq. 5.1) and the graph loss (Eq. 5.2) respectively. Taking all terms into account, the final optimisation objective is,

$$\begin{aligned} \ell(\theta) =& \ell_{sup}(X_S, Y_S; \theta) + \beta_M Tr(\boldsymbol{f}^T \Delta_{\boldsymbol{f}_C} \boldsymbol{f}) \\ & - \beta_C \frac{1}{N} \sum_{i=1}^{N} < \boldsymbol{f}_{C,i}, \log(\boldsymbol{f}_{C,i}) > . \end{aligned} \tag{5.3}$$

where β_C and β_M are balancing hyper-parameters. In order to select β_C and β_M, the model is first run by setting all weights to 1; after the first few iterations, the value of each loss is checked. We then set the two hyper-parameters to rescale the losses to a similar range so that all three terms contribute approximately equally to the training.

Mini-batch Organisation Deep Neural Networks (DNNs) are usually trained with SGD mini-batch optimisation, but Eq. 5.3 is expressed in a full-batch fashion. Converting Eq. 5.3 to mini-batch optimisation is straightforward. However, it is worth mentioning the mini-batch scheduling: each mini-batch contains samples from both source and target domains. The supervised loss

is applied only to source samples with corresponding supervision, the entropy and graph losses are applied to both, and the graph is built per-mini-batch. In this work, the number of source and target samples are equally balanced in a mini-batch.

5.2 Results and Analysis on Common Factorised Space Model

The proposed model is evaluated on multiple visual applications. First, CFSM on unsupervised domain adaptive (UDA) multi-view learning is evaluated (Section 5.2.1). Second, the relevant semi-supervised disjoint label space transfer learning (semi-supervised DLSTL) recognition experiment (Luo et al, 2017) is launched (Section 5.2.2). Moreover, the CFSM copes with the conventional UDA setting (Section 5.2.3). All these scenarios can be handled using CFSM with minor modifications. The effectiveness of CFSM is demonstrated by its superior performance compared to the existing work. Finally, insights are provided through ablation study and visualisation analysis.

5.2.1 Unsupervised Domain Adaptive Multi-View Learning

Person Re-ID The person re-identification (Re-ID) problem is to match person detections across camera views. Annotating person image identities in every camera in a camera network for training supervised models is infeasible. This motivates the topical UDA person Re-ID problem of adapting a Re-ID model trained on one dataset with annotation to a new dataset without annotation. Although they are evaluated with retrieval metrics, contemporary Re-ID models are trained using identity prediction (classification) losses. This means that UDA person Re-ID follows the UDA multi-view setting, as the label spaces (person identities) are different in different Re-ID datasets, and the target dataset has no labels.

Two highly contested large-scale benchmarks for UDA person Re-ID: Market-1501 (Zheng et al, 2015) and DukeMTMC-reID (Zheng et al, 2017), are adopted. ImageNet pre-trained Resnet50 (He et al, 2016a) is used as the feature extractor Φ_M. Cross-entropy loss with label smoothing and triplet loss are used for the source domain as supervised learning objectives. We set $d_C = 2048, \beta_M = 2.0, \beta_C = 0.01$. Adam optimiser is used with learning rate $3.5e^{-4}$. Each dataset in turn is treated as source/target. Rank 1 (R1) accuracy and mean Average Precision (mAP) results on the target datasets are used as evaluation metrics.

In Table 5.1, the proposed CFSM outperforms the state-of-the-art alternatives purpose-designed

	M2D		D2M	
	R1	mAP	R1	mAP
UMDL (Peng et al, 2016)	18.5	7.3	34.5	12.4
PTGAN (Wei et al, 2018)	27.4	-	38.6	-
PUL (Fan et al, 2018b)	30.0	16.4	45.5	20.5
CAMEL (Yu et al, 2017a)	-	-	54.5	26.3
TJ-AIDL (Wang et al, 2018)	44.3	23.0	58.2	26.5
SPGAN (Deng et al, 2018)	46.4	26.2	57.7	26.7
MMFA (Lin et al, 2018)	45.3	24.7	56.7	27.4
CFSM	**49.8**	**27.3**	**61.2**	**28.3**

Table 5.1: Unsupervised domain adaptive (UDA) person Re-ID (%). M2D indicates Market-1501 as the source domain and DukeMTMC-reID as the target one, vice versa. Target Dataset Performance is reported.

for the UDA person Re-ID. Note that TJ-AIDL (Wang et al, 2018) and MMFA (Lin et al, 2018) exploit attribute labels to help alignment and adaptation. The proposed method automatically discovers latent factors with no additional annotation. However, CFSM improves at least 3.0% over TJ-AIDL and MMFA on the R1 accuracy of both settings.

FG-SBIR Fine-grained Sketch Based Image Retrieval (SBIR) focuses on matching a sketch with its corresponding photo (Sangkloy et al, 2016). As demonstrated in (Sangkloy et al, 2016), object category labels play an important role in retrieval performance, so existing studies make a closed world assumption, i.e., all testing categories overlap with training categories. However, if deploying SBIR in a real application such as e-commerce (Yu et al, 2016), one would like to train the SBIR system once on some source object categories, and then deploy it to provide sketch-based image retrieval of new categories without annotating new data and re-training for the target object category. Unsupervised adaptation to new categories without sketch-photo pairing labels is therefore another example of the UDA multi-view learning problem. Comparing to Re-ID, where instances are person images in different camera views, instances in SBIR are either photos or hand-drawn sketches of objects.

There are 125 object classes in the Sketchy dataset (Sangkloy et al, 2016). We randomly split 75 classes as a labelled source domain and use the remaining 50 classes to define an unlabelled target domain with disjoint label space. ImageNet pre-trained Inception-V3 (Szegedy et al, 2016) is used as the feature extractor Φ_M. Cross-entropy and triplet loss are used for source supervision.

We set $d_C = 512, \beta_M = 10^{-3}, \beta_C = 0.1$. Adam optimiser with learning rate 10^{-4} is used. As a baseline, Source Only is the direct transfer alternative that uses the same architecture but trains on the source labelled data only, and is applied directly to the target without adaptation. The retrieval performance on unseen classes (tar. cls.) are reported. Results are averaged over 10 random splits. As shown in Table 5.2, the proposed CFSM improves the retrieval accuracy on unseen cases by 2.48%.

	Source only	CFSM
tar. cls.	23.74 ± 0.24	$\mathbf{26.22 \pm 0.25}$

Table 5.2: SBIR: Sketch-photo retrieval results (%). Averaged Rank 1 accuracy and standard error.

5.2.2 Semi-supervised Disjoint Label Space Transfer Learning

Dataset and Settings We follow the semi-supervised DLSTL recognition experiment of (Luo et al, 2017) where again two digit datasets, SVHN and MNIST, are used. Images of digits 0 to 4 from SVHN are fully labelled as source data while images of digits 5 to 9 from MNIST are target data. The target dataset has sparse labels (l labels per class) and unlabelled images available. Thus a classifier χ_T after the CFS layer Φ_C is also added for the target categories.

The feature extractor architecture Φ_M is exactly the same as in (Luo et al, 2017) for fair comparison. CFSM on source data is pre-trained as initialisation, and then train it with both source and target data using only loss in Eq. 5.3. We set $d_C = 10, \beta_M = \beta_C = 0.01$. The learning rate is 0.001 and the Adam (Kingma and Ba, 2014) optimiser is used.

Results The results for several degrees of target label sparsity $l = 2,3,4,5$ (corresponding to $10,15,20,25$ labelled samples, or $0.034\%, 0.050\%, 0.066\%, 0.086\%$ of total target training data respectively), are reported in Table 5.3. Results are averaged over ten random splits as in (Luo et al, 2017). Besides the FT matching nets (Vinyals et al, 2016) and state-of-the-art LET results from (Luo et al, 2017), two baselines are run: Train Target: Training CFSM architecture from scratch with partially labelled target data only, and FT Target: The standard pre-train/fine-tune pipeline, i.e., pre-train on the labelled source, and fine-tune on the labelled target samples only.

As shown in Table 5.3, the performances of baseline models are significantly lower than LET and the proposed CFSM. The Train Target baseline performs poorly as it is hard to achieve good performance with few target samples and no knowledge transfer from source. The Fine-

	$l=2$	$l=3$	$l=4$	$l=5$
Train Target	66.5 ± 1.7	77.2 ± 1.1	83.0 ± 0.9	88.3 ± 1.1
FT Target	69.8 ± 1.6	79.1 ± 1.2	84.5 ± 0.8	89.3 ± 0.9
FT matching nets (Vinyals et al, 2016)	64.5 ± 1.9	75.5 ± 2.4	79.3 ± 1.3	82.7 ± 1.1
LET (Luo et al, 2017)	91.7 ± 0.7	93.6 ± 0.6	94.2 ± 0.6	95.0 ± 0.4
CFSM	$\mathbf{93.5 \pm 0.5}$	$\mathbf{94.8 \pm 0.5}$	$\mathbf{95.5 \pm 0.3}$	$\mathbf{96.7 \pm 0.2}$

Table 5.3: Semi-supervised DLSTL image categorisation results (%), with mean classification accuracy and standard error for SVHN (0-4) $\rightarrow$ MNIST (5-9).

Tune Target baseline performs poorly as the annotation here is too sparse for effective fine-tuning on the target problem. Fine-Tune matching nets follows the 5-way $(l-1)$-shot learning with sparsely labelled target data only, but no improvement is shown over the other baselines. Our proposed CFSM consistently outperforms the state-of-the-art LET alternative. For example, under the most challenging setting ($l=2$), CFSM is 1.8% higher than LET on mean accuracy and 0.2% lower on standard error.

5.2.3 Unsupervised Domain Adaptation

Dataset and Settings We evaluate the conventional UDA setting from (Ganin et al, 2016) where SVHN (Netzer et al, 2011) is the labelled source dataset and MNIST (LeCun et al, 1998) is the unlabelled target. For fair comparison an identical feature extractor network is use to (Luo et al, 2017). Our CFSM is pre-trained on the source dataset with cross-entropy supervision and $d_C = 50$, followed by joint training on source and target with our regularisers as in Eq. 5.3. Since the label space is shared in UDA, entropy loss is also applied on the softmax classification of the target (Long et al, 2016). $\beta_M = 0.001$ and $\beta_C = 0.01$ are set.

Results Our method is compared with two baselines. Source only: Supervised training on the source and directly apply to target data. Joint FT: Model is initialised with source pre-train, and fine-tuning on both domains with supervised loss for source and semi-supervised entropy loss for target.

As shown in Table 5.4, CFSM boosts the performance on both baselines with clear margin (25.5% and 9.3% v.s. Source only and Joint FT respectively). Moreover, it is 5.5% higher than LET (Luo et al, 2017), the nearest competitor and only alternative that *also* addresses the DLSTL setting.

Method	Accuracy
Domain Confusion (Tzeng et al, 2015)	68.1
Gradient Reversal (Ganin et al, 2016)	73.9
ADDA (Tzeng et al, 2017)	76.0
LET (Luo et al, 2017)	81.0
Res-para (Rozantsev et al, 2018)	84.7
Asym. tri-train (Saito et al, 2017)	85.0
Source only	61.0
Joint FT	77.2
CFSM	**86.5**

Table 5.4: Unsupervised domain adaptation results. Classification accuracy (%) on SVHN→MNIST transfer.

5.2.4 Further Analysis

Ablation Study Unsupervised domain adaptive (UDA) person Re-ID is chosen as the main benchmark for an ablation study. Firstly because it is a challenging and realistic large-scale problem in the UDA multi-view setting, and secondly because it provides a bidirectional evaluation for more comprehensive analysis.

The following ablated variants are proposed and compared with the full CFSM. Source Only: The proposed architecture is learned with source data and supervised losses only. Source+Regs: The regularisers, unsupervised factorisation and graph losses can be added with source dataset only. CFSM−Graph: Our method without the proposed graph loss. CFSM+ClassicGraph: Replacing our proposed graph loss with a conventional Laplacian graph (i.e., graphs constructed in lower-level feature space extracted by Φ_M to regularise the proposed CFS). AE: Other regularisers such as feature reconstruction as in autoencoder (AE) is used to provide the prior term $p(\theta|X)$. The deep features $\boldsymbol{f}$ are reconstructed using the outputs of CFS layer as hidden representations. In this case both source and target data are used and the reconstruction error provides the regularisation loss. The results are shown in Table 5.5. Firstly, by comparing the variants that use source data only (Source Only and Source+Regs) with the joint training methods, they are consistently inferior. This illustrates that it is crucial to leverage target domain data for adaptation. Secondly, CFSM and its variants consistently achieve better results than AE, illustrating that our unsupervised factorisation loss and graph losses provide better regularisation for cross-domain/cross-task adaptation. The effectiveness of our graph loss is illustrated by two compar-

	M2D		D2M	
	R1	mAP	R1	mAP
Source Only	39.2	20.2	54.4	23.0
Source+Regs	41.6	21.2	55.8	24.0
AE	43.6	22.8	56.4	24.9
CFSM−Graph	46.8	25.6	60.0	27.6
CFSM+ClassicGraph	47.4	26.1	59.0	27.0
CFSM	**49.8**	**27.3**	**61.2**	**28.3**

Table 5.5: Ablation study on UDA person Re-ID. Target performance (%) is reported.

isons: (1) CFSM−Graph is worse than CFSM, showing the contribution of the graph loss; and (2) replacing our graph loss with the conventional Laplacian graph loss (CFSM+ClassicGraph) shows worse results than ours, justifying our choice of regularisation direction. Finally, applying our regularisers to the source only (Source+Regs) still improves the performance slightly on target dataset vs Source Only.

Visualisation Analysis To understand the impact of unsupervised factorisation loss, Figure 5.2 illustrates the distribution of target CFS activations in the semi-supervised DLSTL setting (SVHN $\rightarrow$ MNIST). The left plot shows the activations without any such loss, leading to a distribution of moderate predictions peaked around 0.5. In contrast, the right plot shows the activation distribution on the target dataset of CFSM. It is shown that our regulariser has indeed induced the target dataset to represent images with a low–entropy near-binary code. We also compare training a source model by adding low-entropy CFS loss, and then applying it to the target data. This leads to a low-entropy representation of the source data, but the middle plot shows that when transferred to the target dataset or adaptation the representation becomes high-entropy. That is, joint training with our losses is crucial to drive the adaptation that allows target dataset to be represented with near-binary latent factor codes.

Qualitative Analysis The discovered latent attributes are qualitatively visualised. For each element in $\boldsymbol{f}_C$, images in both source and target domains are ranked by their activation. Person images corresponding to the highest ten values of a specific $\boldsymbol{f}_C$ are recorded. Figure 5.3 shows two example factors with images from the source (first row) and target (second row) dataset. We can see that the first example in Figure 5.3(a) is a latent attribute for the colour 'red' covering both people's bags and clothes. The second example in Figure 5.3(b) is a higher-level latent attribute

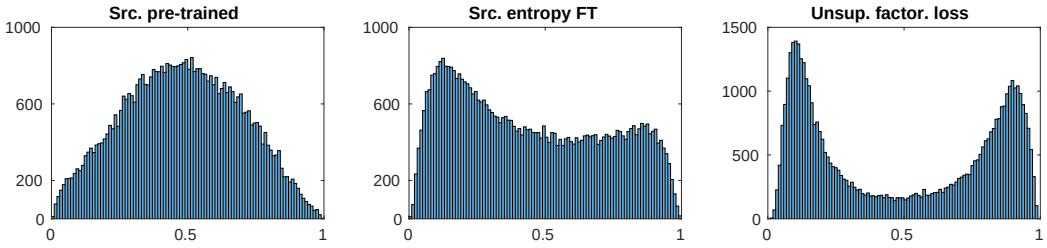


Figure 5.2: CFS activations distribution on target data. Left: Train on source with supervised loss. Middle: Train on source with both supervised and low-entropy CFS losses. Right: CFSM, jointly trained on source and target.

that is selective for both females, as well as textured clothes and bag-carrying. Importantly, these factors have become selective for the same latent factors across datasets, although the target dataset has no supervision (i.e., UDA multi-view learning).

Figure 5.3: Images selected by two latent factors: (a) red and (b) female/textured/bag-carrying. In each case the top row is the source (Market) data and the bottom row is the target (Duke) data.

5.3 Stochastic Inference for Deep Clustering

In this section, a novel unsupervised domain adaptive multi-view learning algorithm is proposed based on learning a joint feature embedding that encourage the target domain data to form clusters. The key component is a new deep clustering method called Stochastic Inference for Deep

Clustering (SIDC) for the target domain. It jointly learns representation and clustering by treating cluster assignment as a random, rather than deterministic, variable in order to alleviate compounding errors at early training. Furthermore, a new triplet loss for target data is also formulated based on the learned cluster centres and stochastic assignments. The proposed method is applied on a challenging UDA multi-view task, the unsupervised domain adaptive (UDA) person Re-ID to be specific, and achieves state-of-the-art performance. Moreover, our SIDC is effective on its own on the conventional data clustering task.

5.3.1 Overview

Problem Setting For unsupervised domain adaptive person Re-ID, we have a labelled source domain S where a dataset contains both person images X_S and their identity labels Y_S; we also assume that for a target domain/dataset, only person images X_T are available without identity labels. The objective is to improve matching performance on the target domain by transferring knowledge from the source domain via a shared feature embedding space learned with both X_S and X_T. Our multi-task joint feature learning model is a deep convolutional neural network (CNN) composed of modules (sub-networks) that are either shared across domains or domain-specific (see Figure 5.4).

Shared Modules A CNN feature extractor Φ_M is used to extract the appearance representation $\boldsymbol{f} \in \mathbb{R}^{d_f}$ of an input person image x. Φ_M is shared across source and target domains, and $\boldsymbol{f} = \Phi_M(x)$ will also serve as the final feature for cross-camera matching at testing time. An encoding layer Φ_{enc} then encodes $\boldsymbol{f}$ into a hidden feature $\boldsymbol{h} \in \mathbb{R}^{d_h}$, with $\boldsymbol{h} = \Phi_{enc}(\boldsymbol{f})$, before a decoding network Φ_{dec} is used to reconstruct the CNN feature, that is, $\hat{\boldsymbol{f}} = \Phi_{dec}(\Phi_{enc}(\boldsymbol{f})) \approx \boldsymbol{f}$.

Domain-Specific Modules To make use of the source dataset labels for feature learning, a person identity classifier $\Phi_{cls}(\boldsymbol{h})$ is applied to the hidden encoding $\boldsymbol{h}$ to predict person identities Y_S for source data X_S only. Via back-propagation, this ensures that the feature extractor Φ_M and encoder Φ_{enc} combined learn a feature representation that is discriminative for the source domain identities. More importantly, this representation must be effective for the target data X_T. To that end, our Stochastic Inference for Deep Clustering (SIDC) is formulated on the hidden representation $\boldsymbol{h}$ for the target domain data only. This introduces no new parameters besides the set of k cluster centres $\boldsymbol{\psi} = [\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k]$, $\boldsymbol{\psi} \in \mathbb{R}^{d_h \times k}$. Moreover, a triplet loss based on the tuples composed of the clusters $\boldsymbol{\psi}$ and target training samples is employed. Our model is explained in detail in Section 5.3.2.

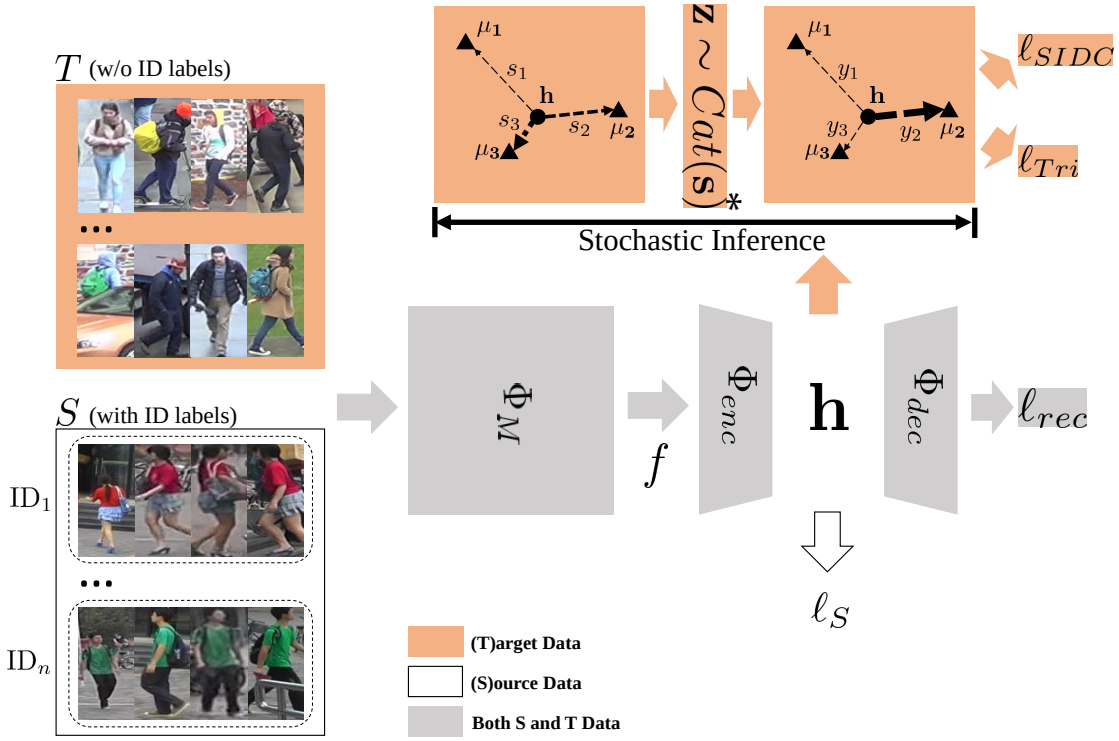


Figure 5.4: Illustration of the proposed multi-task joint feature learning framework for UDA person Re-ID. '$*$' indicates the reparameterisation trick is applied in order to backpropagate through the stochastic sampling $\boldsymbol{z} \sim \text{Cat}(\boldsymbol{s})$.

Summary The proposed model is trained on both source and target data in a multi-task learning procedure with four losses: a supervised loss ℓ_S on the source domain, both the unsupervised SIDC loss ℓ_{SIDC} for clustering and triplet loss ℓ_{Tri} are based on the stochastic inference of unlabelled X_T, and a standard unsupervised autoencoder reconstruction loss ℓ_{rec} that is applied on both domains to regularise the clustering. θ is used to summarise the trainable parameters in Φ_M, Φ_{enc}, Φ_{dec} and Φ_{cls}, and $\boldsymbol{\psi}$ to denote the new cluster centre parameters unique to the novel SIDC component. The overall loss is:

$$L_{\theta,\boldsymbol{\psi}} = \ell_S(X_S,Y_S) + \alpha\ell_{SIDC}(X_T) + \gamma\ell_{Tri}(X_T) + \beta\ell_{rec}(X_S,X_T), \tag{5.4}$$

where α, γ and β are weighting parameters. For the supervised source-domain specific loss ℓ_S, both cross entropy with label smoothing (Szegedy et al, 2016) and mini-batch hard triplet loss (Hermans et al, 2017) are used. For mini-batch construction, each mini-batch contains an equal number of samples from both datasets. The whole pipeline is illustrated in Figure 5.4.

The intuition here is that we want to train a feature extractor $\boldsymbol{f} = \Phi_M(\cdot)$ that simultaneously: (i) supports accurate person Re-ID in the source domain where identities are known (ℓ_S), and also (ii) is adapted to the target domain in the sense of cleanly grouping people into clusters

(ℓ_{SIDC}). Each of the learned clusters ideally should contain one unique identity. In practice, however since no assumption is made on the number of identities in X_T, we only assume that the discovered clusters loosely correspond to the actual person identities. It should show that simply requiring the target domain data to support a good clustering structure provides strong cues for unsupervised domain adaptation.

5.3.2 Stochastic Inference for Deep Clustering

Loss Formulation An existing approach to deep clustering like (Yang et al, 2017a) would (using our notation) use ℓ_{rec} and ℓ_{SIDC} to define an optimisation where the loss is:

$$\begin{gathered}\min_{\theta,\boldsymbol{\psi},\boldsymbol{z}} ||\Phi_{enc}(\boldsymbol{f}) - \boldsymbol{\psi}\boldsymbol{z}|| + ||\boldsymbol{f} - \Phi_{dec}(\Phi_{enc}(\boldsymbol{f}))|| \\ s.t.\ \mathbf{1}^T\boldsymbol{z} = 1, \boldsymbol{z} \in \{0,1\}^k\end{gathered} \tag{5.5}$$

where $\boldsymbol{z}$ is a hard cluster indicator and the reconstruction term prevents trivial solutions[1]. However, this cannot be trained jointly end-to-end because hard assignment of clusters prevents back-propagation into the feature extractor – hence the alternating optimisation proposed by (Yang et al, 2017a). It also suffers from compounding errors for the same reason. Our SIDC provides an end-to-end trainable solution for clustering that is resistant to reinforcing errors.

Our approach is built upon a probabilistic clustering assumption where a data point is *more likely* to belong to a closer centre. Therefore, the assignment of each data point $\boldsymbol{h}$ (in the autoencoder space defined previously) to a cluster is treated as a random variable that will be inferred with a stochastic sampling process.

Firstly, a probability vector $\boldsymbol{s}(\boldsymbol{h},\boldsymbol{\psi}) = [s_1,\ldots,s_k]$ is defined to quantify the probability of association, or soft assignment, of the data point $\boldsymbol{h}$ to each of the k clusters.

$$s_i = \frac{\exp(-d(\boldsymbol{h},\boldsymbol{\mu}_i))}{\sum_{j=1}^{k}\exp(-d(\boldsymbol{h},\boldsymbol{\mu}_j))}, \tag{5.6}$$

where $d(\boldsymbol{h},\boldsymbol{\mu}_i)$ is the squared Euclidean distance between $\boldsymbol{h}$ and cluster centre $\boldsymbol{\mu}_i$.

In the forward pass, this distribution is *sampled* in order to stochastically assign $\boldsymbol{h}$ to a cluster:

$$\boldsymbol{z} \sim \mathrm{Cat}(\boldsymbol{s}), \tag{5.7}$$

where $\mathrm{Cat}(\boldsymbol{s})$ indicates the categorical distribution parameterised by $\boldsymbol{s}$ and $\boldsymbol{z}$ is a one-hot sample from $\boldsymbol{s}$.

[1]Typical trivial solutions include that all $\boldsymbol{h}$ become zero vectors (Yang et al, 2017a) and/or all data points are assigned to one cluster.

This sampling process is used to define a loss that prefers configurations where $\boldsymbol{z}$ and $\boldsymbol{s}$ are more similar. That is, where $\boldsymbol{s}$ tends to be one-hot. This in turn corresponds to situations where clusters are well separated from each other and each data point $\boldsymbol{h}$ is close to its nearest cluster centre. Specifically ℓ_{SIDC} is defined as:

$$\ell_{SIDC} = \mathrm{KL}(\boldsymbol{z}||\boldsymbol{s}) = \sum_{j=1}^{k} z_j \log \frac{z_j}{s_j}. \tag{5.8}$$

If the learned representation $\boldsymbol{h}$ comes from a *clustering-friendly* (Yang et al, 2017a) space where target data naturally forms tight and well-separable clusters, then $\boldsymbol{s}$ (Eq. 5.6) can clearly become close to one-hot. While if the space is not clustering-friendly, $\boldsymbol{s}$ will be far from one-hot. Thus training the feature extractor $\Phi_M(\cdot)$ and encoder $\Phi_{enc}(\cdot)$ to optimise ℓ_{SIDC} will promote a cluster friendly embedding space $\boldsymbol{h}$.

Note that unlike deterministic approaches such as (Xie et al, 2016a), which iteratively increase the strength of soft predictions to match the current most likely prediction, our SIDC approach pulls $\mathrm{Cat}(\boldsymbol{s})$ in different directions on different iterations due to the stochastic sampling $\boldsymbol{z} \sim \mathrm{Cat}(\boldsymbol{s})$. This randomness helps the training escape from situations of reinforcing error, by exploring the different possible assignments – especially for low-confidence points (see Figure 5.6 for a visualisation).

A Differentiable Approximation The problem with using the loss defined in Eq. 5.8 as objective is that the gradients cannot be back-propagated through the stochastic sampling layer (Eq. 5.7). To overcome this problem, the Gumbel-Softmax reparameterisation trick (Jang et al, 2017; Maddison et al, 2017) is employed. As the replacement of z_i, its differentiable approximation y_i is used, defined as,

$$y_i = \frac{\exp((g_i + \log s_i)/\tau)}{\sum_{j=1}^{k} \exp((g_j + \log s_j)/\tau)} \tag{5.9}$$

where g_i is i.i.d sampled from $\mathrm{Gumbel}(0,1)$ distribution[2] and τ is a temperature factor that tunes how closely the softmax function approximates the non-differential argmax function. $\tau = 0.1$ for all experiments. Using Eq. 5.9, the ℓ_{SIDC} loss in Eq. 5.8 is expressed as

$$\ell_{SIDC} = \mathrm{KL}(\boldsymbol{y}||\boldsymbol{s}) = \sum_{j=1}^{k} y_j \log \frac{y_j}{s_j}. \tag{5.10}$$

Triplet Loss for Unlabelled Target Data A target sample $\boldsymbol{h}_{t_p}$ is assigned to a cluster centre $\boldsymbol{\mu}_{z_{t_p}}, z_{t_p} \in \{1,...,k\}$ in $\boldsymbol{\psi}$ based on the stochastic inference (Eq. 5.7) and they form a positive pair

[2]To sample $\mathrm{Gumbel}(0,1)$ distribution, a random number r is draw from the uniform distribution $[0,1]$ and then compute $g = -\log(-\log(r))$.

$\{\boldsymbol{\mu}_{z_{t_p}}, \boldsymbol{h}_{t_p}\}$. The cluster centre $\boldsymbol{\mu}_{z_{t_p}}$ is then used as an anchor to retrieve the hard negative (Hermans et al, 2017) target sample $\boldsymbol{h}_{t_n}$ within each mini-batch, where $t_n = \text{argmin}_t\, d(\boldsymbol{\mu}_{z_{t_p}}, \boldsymbol{h}_t)$, $s.t.\ z_t \neq z_{t_p}$. And the triplet loss for the tuple $(\boldsymbol{\mu}_{z_{t_p}}, \boldsymbol{h}_{t_p}, \boldsymbol{h}_{t_n})$ is,

$$\ell_{Tri} = [m + d(\boldsymbol{\mu}_{z_{t_p}}, \boldsymbol{h}_{t_p}) - d(\boldsymbol{\mu}_{z_{t_p}}, \boldsymbol{h}_{t_n})]_+ \tag{5.11}$$

where $m \geq 0$ is the margin value. This clustering-specific triplet loss is designed to further make each cluster compact and farther away from other clusters.

Reconstruction Loss With the clustering loss alone, (i.e., Eq. 5.10), deep models may produce trivial solutions as discussed in (Yang et al, 2017a). To alleviate this problem, clustering loss should be combined with a reconstruction loss. The loss ℓ_{rec} is applied to both source and target Re-ID datasets,

$$\ell_{rec} = ||\boldsymbol{f} - \Phi_{dec}(\Phi_{enc}(\boldsymbol{f}))||_1, \tag{5.12}$$

which is a L_1 loss.

5.4 Results and Analysis on Stochastic Inference for Deep Clustering

5.4.1 Experimental Settings

Datasets Three large-scale person Re-ID benchmarks, Market-1501 (Zheng et al, 2015), CUHK03 (Li et al, 2014) and DukeMTMC-reID (Zheng et al, 2017) are used for UDA Re-ID evaluation. **Market-1501** has $12,936$ training and $19,732$ testing images of $1,501$ different identities captured by 6 cameras. We use the standard train/test split in (Zheng et al, 2015) with 751 PIDs in the training set and the remaining 750 PIDs for testing. **DukeMTMC-reID** is a subset of the Duke dataset (Ristani et al, 2016) for person Re-ID purposes. We follow the protocol in (Zheng et al, 2017) where $16,522$ images of 702 PIDs are used as the training set, and the other 702 PIDs are in testing set. **CUHK03** contains $14,096$ images of $1,467$ PIDs, where 767 PIDs are in the training set and the remaining 700 people are in the testing set. The detected person bounding boxes are used following (Zhong et al, 2017). Two UDA Re-ID settings have been used in existing studies; both of them are thus adopted. **Setting 1:** Market-1501 and DukeMTMC-reID are used, one as source and the other as target, resulting in two directions of transfer. **Setting 2:** all three benchmarks are used with one as source and the other two as target, giving six transfer directions in total. So Setting 1 is a subset of Setting 2. Under both settings, the training splits of both the source and target (unlabelled) are used for training and the test split of the target for testing.

Data Augmentation The input person images are all fixed at size 256×128. Random horizontal flips are used during training. The mini-batch size is 64 with half of the images from source and the others from target.

Model Architecture Similar to most existing UDA person Re-ID models, a ResNet-50 based backbone is used. Specifically, ResNet-50 with instance normalisation (Pan et al, 2018) is used as the feature extractor Φ_M for the appearance feature $\boldsymbol{f} \in \mathbb{R}^{2048}$. The encoder Φ_{enc} is a fully connected layer projecting $\boldsymbol{f}$ into the hidden space $\boldsymbol{h} \in \mathbb{R}^{512}$. The decoder Φ_{dec} then decodes $\boldsymbol{h}$ to reconstruct $\boldsymbol{f}$ with a 1024D intermediate layer. The PID classifier Φ_{cls} for source dataset projects $\boldsymbol{h}$ into the label space. $k = 1,200$ clusters are used as default for all UDA person Re-ID experiments, regardless what the true target training identity numbers are. This is unlike PUL (Fan et al, 2018b) which sets k to be almost identical to the true identity number in each target dataset. The impacts of different ks on performance are also studied (see Sec. 5.4.3).

Optimisation During pre-training, the mapping network Φ_M is initialised with ImageNet-pretrained weights, and then fine-tuned. Other components are trained from scratch. No clustering is applied yet at this stage (i.e., α and γ are 0) and β is fixed at 0.1 in Eq. 5.4. Initial learning rate is 3.5×10^{-4}. Adam (Kingma and Ba, 2014) optimiser is used and trained with $80,000$ iterations. Our clustering loss ℓ_{SIDC} and triplet loss ℓ_{Tri} are then applied to the target data and jointly fine-tuned along with the rest of the model and other losses. Loss balancing hyper-parameters α, γ and β are fixed at 0.01, 0.01 and 1.0 respectively. The margin m of triplet loss ℓ_{Tri} is fixed at 0.2. Learning rate is set to 1.0×10^{-4} with $50,000$ training iterations. The sensitivity of model against the hyper-parameters is also studied (see Sec. 5.4.3).

Evaluation Metrics The top Ranked matching accuracy (Rank 1) and mean average precision (mAP) are used.

5.4.2 Results

Results under Setting 1 Table 5.6 compares our model with existing state-of-the-art alternatives under Setting 1. The following specific observations are made: (1) Our model beats the existing models by a significant margin under both transfer directions, e.g., 12.3% on mAP over the nearest competitor CFSM (Section 5.1) for M→D. (2) Unsurprisingly, the hand-crafted feature based UMDL (Peng et al, 2016) is the weakest. (3) TJ-AIDL (Wang et al, 2018) and MMFA (Lin et al, 2018) are the two feature alignment based methods. Despite both requiring additional attribute labels from the source domain, their performance is significantly lower than

	M→D		D→M	
	R1	mAP	R1	mAP
UMDL (Peng et al, 2016)	18.5	7.3	34.5	12.4
PTGAN (Wei et al, 2018)	27.4	-	38.6	-
PUL (Fan et al, 2018b)	30.0	16.4	45.5	20.5
TJ-AIDL (Wang et al, 2018)	44.3	23.0	58.2	26.5
SPGAN (Deng et al, 2018)	46.4	26.2	57.7	26.7
MMFA (Lin et al, 2018)	45.3	24.7	56.7	27.4
HHL (Zhong et al, 2018)	46.9	27.2	62.2	31.4
CFSM (Section 5.1)	49.8	27.3	61.2	28.3
SIDC	**61.3**	**39.6**	**69.2**	**37.9**

Table 5.6: UDA Re-ID results (%) under Setting 1. M→D indicates Market-1501 as source dataset and DukeMTMC-reID as target, and vice versa. Target dataset performance is reported.

ours. (4) Among the three models that employ GAN-based image synthesis, PTGAN (Wei et al, 2018) and SPGAN (Deng et al, 2018) attempt to transfer the image style from source to target. The results show that this cross-domain style alignment is clearly inferior to the latest image synthesis-based method HHL (Zhong et al, 2018), which synthesises images across target camera views (sub-domains) only. Importantly, like our model, HHL does source-target joint feature learning. However, our model, without needing the tricky GAN training and using a simple clustering learning objective, is superior to all three GAN-based models. (5) Both PUL (Fan et al, 2018b) and our model follow a clustering paradigm for UDA person Re-ID. However, the proposed SIDC significantly improves the performance over PUL (Fan et al, 2018b). Without multi-task learning and relying on hard deterministic cluster assignment, PUL is clearly ineffective. (6) The best results reported so far are obtained by CFSM, which also benefits from joint feature learning. However there is still a massive gap between its performance and ours, indicating that discovering identities by clustering is more effective than discovering latent attributes.

Results under Setting 2 As shown in Table 5.7, only three competitors reported results under this setting. Since Setting 1 is a subset of Setting 2, only the transfer directions not in Setting 1 are reported in Table 5.7. Again, our model outperforms the compared models significantly under all directions. For instance, compared with the nearest competitor HHL (Zhong et al, 2018), under C→M the proposed SIDC improves the mAP by 16.4%.

	C→M		C→D		M→C		D→C	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
PUL	41.9	18.0	23.0	12.0	7.6	7.3	5.6	5.2
SPGAN	42.3	19.0	-	-	-	-	-	-
HHL	56.8†	29.8†	42.7†	23.4†	-	-	-	-
SIDC	**71.2/74.4**†	**41.9/46.2**†	**45.5/47.6**†	**25.2/28.5**†	**22.9**	**24.0**	**21.0**	**22.5**

Table 5.7: UDA Re-ID results (%) under Setting 2. C, M, and D refer to CUHK03, Market-1501 and DukeMTMC-reID respectively. † indicates the HHL protocol whereby both CUHK03 training and test splits are used as source data.

Ablation Study Our model is based on multi-task feature learning with the novel SIDC loss ℓ_{SIDC} serving for the stochastic clustering objective with probabilistic cluster assignment. The learned cluster centres $\boldsymbol{\psi}$ enable the proposed clustering-specific triplet loss ℓ_{Tri} for the target domain and a reconstruction loss ℓ_{rec} for both domains. In this experiment, their contributions are evaluated to our final model performance. A few variants of our full model are considered. **Direct Transfer (DT):** without multi-task learning, only the source dataset is used for learning a feature extractor, which is then used directly for the target domain. **Auto-encoder Transfer (AE Transfer):** now we add ℓ_{rec} and remove ℓ_{SIDC} and ℓ_{Tri}, i.e., the unlabelled target data is still used during training, but instead of for clustering, for self-reconstruction purpose only. **DCN (Yang et al, 2017a):** The deep model (including ℓ_{rec}), and cluster centres and cluster assignments are optimised alternatively based on hard deterministic inference. **DEC (Xie et al, 2016a) + AE:** ℓ_{SIDC} is replaced with the soft deterministic clustering objective DEC (Xie et al, 2016a) which enables end-to-end training. AE architecture (ℓ_{rec}) is used and without ℓ_{Tri}. **SIDC w/o ℓ_{Tri}:** Our full model with the triplet loss ℓ_{Tri} is disabled. **SIDC:** the full model as in Eq. 5.4. Table 5.8 shows the gap between ours and DT is massive, demonstrating that unsupervised transfer learning using the target data is crucial. Importantly, all components in our model contribute to the improvement. Specifically, AE Transfer with reconstruction loss ℓ_{rec} improves performance by jointly learning with the source and target data. DCN (Yang et al, 2017a) yields the worst results among the deep clustering methods compared due to its hard deterministic inference and alternating optimisation. DEC (Xie et al, 2016a) is better thanks to the soft deterministic inference and end-to-end training. However, with stochastic inference for clustering (ℓ_{SIDC}), the model (*SIDC w/o* ℓ_{Tri}) boosts the results significantly. It thus shows clearly that our SIDC loss is superior to the existing deterministic clustering losses. Moreover, the triplet loss ℓ_{Tri} further boosts the

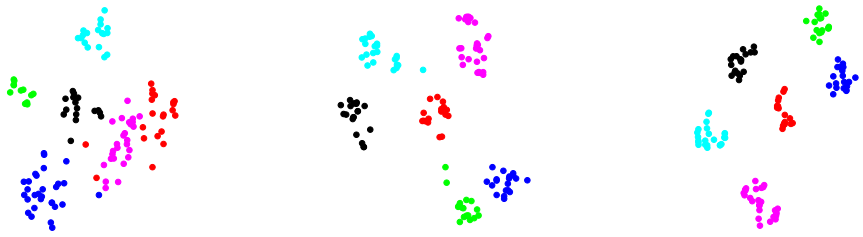
	M→D		D→M		C→M		C→D		M→C		D→C	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DT	40.5	22.1	49.3	21.6	54.3	27.6	31.5	15.0	9.8	11.1	8.7	7.8
AE Transfer	50.0	28.9	61.7	29.1	60.9	30.0	36.7	18.1	10.9	11.7	8.6	10.0
DCN	52.0	30.6	59.1	27.7	62.5	31.9	35.7	17.5	14.1	15.6	11.4	12.4
DEC + AE	53.2	30.8	63.5	31.8	64.3	34.2	37.0	18.7	14.3	15.2	12.9	13.2
SIDC w/o ℓ_{Tri}	60.1	38.5	68.2	36.7	70.0	39.8	42.4	22.7	21.4	22.4	19.5	20.7
SIDC	**61.3**	**39.6**	**69.2**	**37.9**	**71.2**	**41.9**	**45.5**	**25.2**	**22.9**	**24.0**	**21.0**	**22.5**

Table 5.8: Ablation study on UDA person Re-ID under Setting 2. C, M and D refer to CUHK03, Market-1501, and DukeMTMC-reID respectively.

performance by $1 \sim 3\%$ (*SIDC* v.s. *SIDC w/o* ℓ_{Tri}).

5.4.3 Further Analysis

Effectiveness of Stochastic Inference Table 5.8 above shows quantitatively that the proposed clustering loss ℓ_{SIDC} plays a key role in our model and it clearly outperforms the deterministic counterpart DEC (Xie et al, 2016a). We hypothesise that our SIDC loss can better group the target data into clusters than DEC (Xie et al, 2016a) because of the stochastic inference. To validate this, we first show some qualitative results in Figure 5.5. It can be seen clearly that, compared with the DT baseline, test images belonging to different target identities become much more separable in the multi-task feature space $\boldsymbol{f}$ after introducing either clustering loss. In addition, it is clear that different identities become more distinguishable using our SIDC loss, explaining the superior performance in Table 5.8.



(a) DT (b) DEC (c) SIDC

Figure 5.5: t-SNE map of deep features $\boldsymbol{f}$. Different colours represent different PIDs. With Market-1501 as source, (a) shows the target (DukeMTMC-reID) samples in the direct transfer (DT) setting. (b) shows the target features extracted from DEC (Xie et al, 2016a) + AE model and (c) are from SIDC (w/o ℓ_{Tri} for fair comparison) under the M→D setting. Best viewed in colour.

To further illustrate that the better feature learning is indeed caused by better clustering, the embedding space $\boldsymbol{h}$ learned by these two clustering losses are compared, as shown in Figure 5.6. It shows that the initial clustering provided after the first stage of model training using only the supervised loss ℓ_S and the reconstruction loss ℓ_{rec} (see optimisation details in Sec. 5.4.1) is noisy (different identities are assigned to the same clusters). With stochastic inference, SIDC is able to progressively recover from the wrong clustering assignment and form clean clusters. In contrast, even with the softened deterministic assignment, DEC got stuck with the bad initialisation and never recovered.

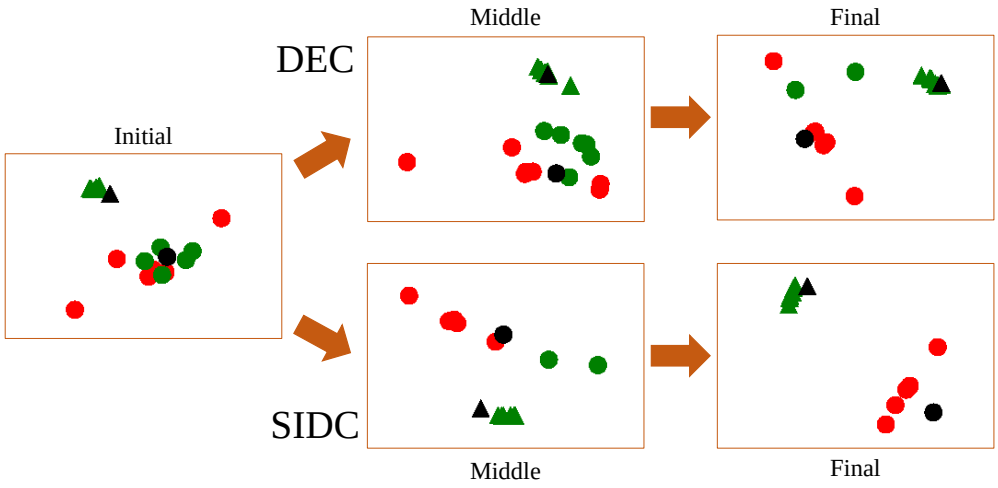


Figure 5.6: An illustration of the clusters formed by DEC (Xie et al, 2016a) + AE and our SIDC at different training stages. Different shapes indicate different cluster memberships, while different colours indicate different person identities. The cluster centres are depicted in black. Two identities (red and green respectively) from DukeMTMC-reID under the M→D setting are shown here in 2D obtained by t-SNE.

Impact of Cluster Number In all experiments so far, the number of clusters are empirically set and fixed at $1,200$, without any dataset-specific tuning. The impact of varying the number of cluster centres k is evaluated in this experiment under the M→D setting. Here the DukeMTMC-reID dataset is the target, whose training split has 702 person identities. Figure 5.7 shows that the performance on the target test set peaks at $k = 1000$. Interestingly setting k to the true identity number does not yield the best result. On the one hand, when k is much smaller than the ground-truth PID number, the performance degradation is severe. On the other hand, the performance becomes stable when k is much bigger. This is explainable: when k is too small, inevitably different identities will be assigned to the same clustering which will have a detrimental effect on the learned feature space. It is also not ideal when k is too big as the same identities will

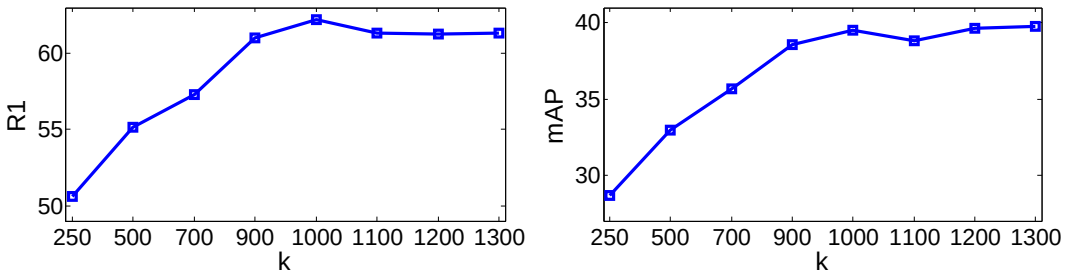


Figure 5.7: Impact of the cluster number k on our model. k is manually set. M→D setting is used and performance (both Rank1 accuracy and mAP) on DukeMTMC-reID is shown.

be allocated to multiple clusters. The good news is that our model seems to be insensitive to that. This result suggests that in practice, it is advisable to over-estimate identity number in the unlabelled target training set.

Model Robustness on Hyper Parameters There are 4 hyper-parameters in our model, namely the loss weights α, γ, β in Eq. 5.4 and the margin m of triplet loss ℓ_{Tri}. The default setting for all experiments are fixed at $\alpha = 0.01$, $\gamma = 0.01$, $\beta = 1.0$ and $m = 0.2$. Their impact on performance is studied in Figure 5.8. Only one hyper-parameter is varied at a time while the others are fixed to the default values. It is shown that in general SIDC performance is stable against the four hyper-parameters within large value ranges.

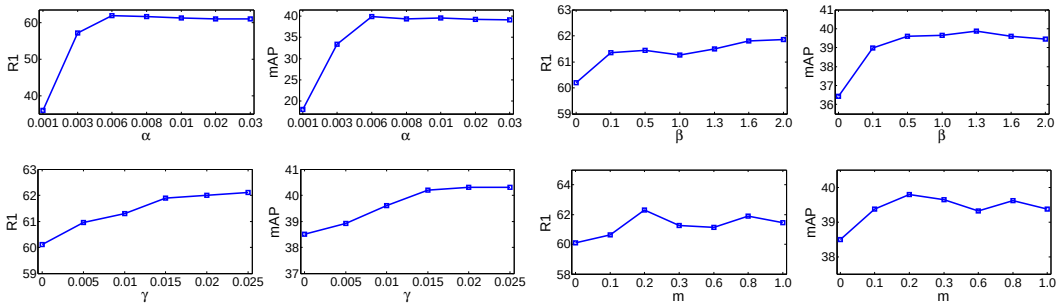


Figure 5.8: Impact of hyper-parameters α (up-left), γ (down-left), β (up-right) and m (down-right). DukeMTMC-reID performance (Rank1 accuracy and mAP) in M→D setting is shown.

SIDC for General Clustering Finally, we evaluate whether the proposed clustering method is also effective when applied on its own to general data clustering tasks. The **MNIST** (LeCun et al, 1998) dataset, which consists of $70,000$ hand-written digit images, is used. The raw pixel data, of size 28×28 in gray scale, are used without pre-processing. No data augmentation is used during training. The mini-batch size is 256. For clustering only, the feature extraction CNN Φ_M is not required. The images are directly fed into the encoder-decoder network (i.e., an autoencoder (AE)). The encoder network is identical to (Xie et al, 2016a; Yang et al, 2017a) which is a 4-

layers MLP with 500, 500, 2000 and 10 hidden neurons. The decoder is mirrored based on the encoder architecture in order to reconstruct the input images. Clustering is performed on the outputs of encoder with $k = 10$. During training, the encoder-decoder is pre-trained from scratch on reconstruction loss only. Initial learning rate is 0.1 and pre-trained with $50,000$ iterations. Adam optimiser is used. Then the model is fine-tuned with both clustering and reconstruction loss. The clustering loss weight is fixed at 0.01. Initial learning rate changes to 0.01 for another $50,000$ iterations training. For clustering evaluation on MNIST, three standard clustering metrics (normalised mutual information (NMI) (Cai et al, 2010a), adjusted random index (ARI) (Yeung and Ruzzo, 2001) and clustering accuracy (ACC) (Cai et al, 2010a)) are used, following most existing works. Higher values are preferred for all metrics.

	NMI	ARI	ACC
k-means	0.50	0.37	0.53
AE + k-means	0.77	0.68	0.78
DEC (Xie et al, 2016a)	0.80	0.75	0.84
DCN (Yang et al, 2017a)	0.81	0.75	0.83
SIDC w/o ℓ_{Tri}	**0.86**	0.78	0.84
SIDC	**0.86**	**0.79**	**0.85**

Table 5.9: Clustering results on MNIST.

The results in Table 5.9 show that our models achieves the best performance. Specifically, we note that: (1) The results of deep models (from AE+k-means to SIDC) are better than shallow ones (k-means). This is expected as vanilla k-means is based on raw image pixels and it cannot modify the feature space. (2) AE+k-means is inferior to other deep models since it does not tune the feature space jointly with the clustering. (3) Although the clustering accuracy (ACC) of SIDC (SIDC w/o ℓ_{Tri}) and DEC (Xie et al, 2016a) are the same, SIDC is 6% and 3% higher than DEC (Xie et al, 2016a) on NMI and ARI respectively. Compared to DCN (Yang et al, 2017a), our models are superior on all three metrics. (4) Our triplet loss ℓ_{Tri} can further improve both ARI and ACC by 1%. These results thus further validate the effectiveness of SIDC due to its stochastic inference based end-to-end training and probabilistic sampling based cluster assignment, which prevent clustering error reinforcement.

5.5 Summary

This chapter aims to handle a challenging multi-view learning task, the unsupervised domain adaptive (UDA) multi-view learning, which has an unlabelled multi-view dataset as target domain and labelled one as source domain. Moreover, the label spaces across domains are disjoint. Therefore, instead of adopting the existing domain alignment methods, the multi-task learning framework is followed. Data from both the source and target domains are exploited for learning a shared deep feature embedding space. The contributions of the source domain come from supervised learning, and the target domain is subjected to different unsupervised losses with specified assumptions. Two different models are proposed based on two distinctive modelling on the target domain.

The first model assumes that the recognition in both domains should be done in a shared latent factor space, considering the label spaces of different datasets are disjoint. An unsupervised factorisation loss is proposed to discover a common set of discriminative latent factors between source and target datasets in a shared embedding space. Therefore, the proposed methods is called common factorised space model (CFSM). To further improve the feature learning for cross-view recognition, the lower level deep feature should be regularised by the higher level latent semantic feature via a novel graph-based loss. Extensive evaluations show that CFSM outperforms a wide range of contemporary techniques on the UDA multi-view learning problems, e.g. the UDA person Re-ID. Moreover, our CFSM is effective on different transfer learning settings, e.g., the semi-supervised DLSTL and the conventional UDA.

The second model utilises the target domain based on the assumption that the unlabelled instances should form clusters. Ideally, the data samples of the same entity/object across views belong to the same cluster. A novel deep clustering method, called Stochastic Inference for Deep Clustering (SIDC), is thus proposed to fulfill such assumption better. The superiority of SIDC over the existing deep clustering methods are in three folds. Firstly, SIDC is based on the stochastic cluster assignment rather than the deterministic one as in existing methods. Therefore, SIDC is more robust to the clustering error reinforcement and premature convergence issues during training. Secondly, the reparameterisation trick is adopted in SIDC and enables the end-to-end joint feature learning and clustering. Finally, a novel triplet loss is proposed to enhance the model performance on the target domain further. It is shown that the proposed method significantly outperforms existing alternatives on the challenging UDA multi-view learning problem, i.e., the

UDA person Re-ID. Moreover, SIDC can be used as a standalone model for general clustering tasks, where it shows promising performance.

The second model SIDC achieves superior performance to the first one CFSM on the UDA multi-view learning task, i.e., UDA person Re-ID, as illustrated in Table 5.6. Both models are based on the multi-task learning framework with similar DNN architectures for learning a shared embedding space. Therefore, such a performance gap mainly comes from the different assumptions they made. Comparing to the factorisation assumption in CFSM, the clustering one in SIDC is more straightforward to reveal the intrinsic structure of the unlabelled multi-view data.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has presented a collection of multi-view learning algorithms for understanding the visual data contains multiple views, i.e., camera angles, modalities and datasets (domains). In particular, a fundamental multi-view learning task, cross-view recognition, for identifying or matching the instances of an entity across different views have been investigated and explored. Different cross-view recognition settings can be described under two criteria, the number of views and the annotations available for the views. We concentrate on three widely used settings: (1) Two-view data with labels. (2) Generalised multi-view data with labels. (3) Unsupervised domain adaptive (UDA) multi-view learning with target multi-view dataset unlabeled. These problems are inherently challenging due to intrinsic visual ambiguities and the significant appearance variations across views. Therefore, the advanced deep neural network (DNN) models are adopted in the proposed methods. Specifically,

1. Chapter 3 **Scalable Deep Canonical Correlation Analysis:** Canonical Correlation Analysis (CCA) objective aims to maximise the correlation of the corresponding instances across views. Existing deep CCA models directly optimise the CCA objective with exact decorrelation, which is computationally expensive and truncate the gradient flows for deep model training. A novel Deep CCA model, Soft CCA, is proposed to overcome these problems. Soft CCA is based on a soft decorrelation version, Stochastic Decorrelation Loss (SDL). SDL enables the CCA objective to be expressed as a loss to be minimised by

gradient descent rather than as an eigen-decomposition problem. As a result, Soft CCA is compatible with the gradient-based end-to-end optimisation and more scalable than existing deep CCA models. The effectiveness of Soft CCA is demonstrated on the supervised two-view datasets. Moreover, SDL applies to various tasks and models and is superior to alternative decorrelation losses.

2. Chapter 4 **Deep Factorisation for Multi-View Learning:** Existing deep multi-view learning methods rely on either deep feature from a single semantic level or additional attribute annotations. In contrast, a novel deep architecture, Multi-Level Factorisation Net (MLFN), can automatically discover the view-invariant latent factors with no attribute labels. For individual input, the discriminative latent factors dynamically identified by different levels of MLFN blocks correspond to specified semantic levels. The higher level latent semantics are modelled by the MLFN block closer to the top. A compact multi-level representation is obtained by aggregating the discriminative semantic information at all levels of MLFN. It can be efficiently fused with the deep feature to complement final representation. Extensive experiments show the superiority of MLFN on not only the supervised multi-view learning but also the objective categorisation.

3. Chapter 5 **Unsupervised Domain Adaptive Multi-View Learning:** Unsupervised domain adaptive (UDA) multi-view learning is a challenging setting with an unlabelled multi-view dataset as the target domain. A supervised multi-view source dataset with the relevant task is incorporated in the model training to improve the target performance. More importantly, the label spaces across domains are assumed to be disjoint. Existing domain alignment methods are based on a different assumption, the shared label space. Therefore, domain alignment methods are not suitable for UDA multi-view learning problems. In this thesis, the multi-task learning framework is adopted for UDA multi-view learning. It aims to learn a shared feature embedding space with domain-specific contributions. Supervised learning is applied to the labelled source data, while different unsupervised losses can be used for the unlabelled target data based on the assumptions made. Specifically, two novel models are proposed with distinctive assumptions.

   The first model assumes there is a shared latent factor space for both domains, and the recognition is performed in this space to guide the discriminative knowledge transfer across domains. Moreover, each latent factor should be interpreted as a discriminative latent at-

tribute. To this end, the latent factor representations from the common space are subjected to an unsupervised factorisation loss to produce the low-entropy, i.e., near binary, codes. Therefore, the proposed method is called common factorised space model (CFSM). Moreover, a novel graph Laplacian-based loss is proposed to encourage the feature-extractor to learn representations that respect the common space manifold from higher-level to improve deep feature learning. The effectiveness of CFSM is demonstrated not only on the UDA multi-view learning but also different transfer learning tasks.

The second model is built on a clustering assumption: the instances from the unlabelled multi-view dataset (target domain) form tight clusters. Each cluster potentially contains the cross-view samples of an object/instance. Specifically, a novel deep clustering method, called Stochastic Inference for Deep Clustering (SIDC), is proposed to serve the clustering purpose. The cluster assignment in SIDC is a stochastic process by sampling from a categorical distribution. It is more robust to the compounding errors that lead to a sub-optimal solution than existing deep clustering based on deterministic assignment. The performance of SIDC also benefit from its end-to-end optimisation which is enabled by the reparameterisation trick. Moreover, a novel triplet loss is derived based on the clustering outcomes to further improve the performance. SIDC achieves the state-of-the-art performance on a challenging UDA multi-view learning problem, i.e., the UDA person Re-ID. It is also an effective method on the conventional clustering task.

These models are originally proposed for several representative multi-view data settings and primarily evaluated on the corresponding cross-view recognition visual understanding tasks. They also have potentials and benefits for dealing with other relevant tasks in computer vision and machine learning, as demonstrated by the experiments. More discussions about the future research directions and work are detailed below.

6.2 Future Work

The potential research directions for future work are summarised as follows to end this thesis. The main concentration is to extend the two models, Multi-Level Factorisation Net (MLFN) from Chapter 4 and Stochastic Inference for Deep Clustering (SIDC) from Chapter 5.

MLFN is capable of automatically discovering and dynamically identifying the discriminative latent factors appeared in each visual input. Such learned latent factors can be treated as

latent attributes. MLFN also provides a compact feature to encode such factors from all levels. Therefore, a comprehensive latent attribute representation is acquired. Comparing with the manually annotated attributes, the latent attribute representation from MLFN can be obtained at a low price since the training and inference of MLFN require no attribute label. Moreover, the manually labelled attributes are biased to the describable visual factors while the MLFN latent attribute representation encodes the discriminative characteristics from multiple levels. They can be complementary to each other. Therefore, the MLFN latent attribute representation can be used as an alternative or a complement to the expensive attribute labels for a wide range of computer vision tasks such as zero-shot learning (Lampert et al, 2013; Romera-Paredes and Torr, 2015) and semantic human parsing (Takagi et al, 2017; Xiao Wang, 2019).

SIDC is an effective model for UDA multi-view learning problems based on a simple and straightforward clustering assumption on the unlabelled multi-view dataset. Two lines of extension work can be considered. One is to concern the more challenging unsupervised multi-view learning setting with an unlabelled multi-view dataset only, e.g., unsupervised person Re-ID (Li et al, 2018a). SIDC is also applicable to these problems since its clustering assumption is still reasonable under the new setting. However, without the supervision from the source domain, the more advanced deep clustering models are thus required to compensate. To this end, the other concern is to develop the SIDC further. One limitation of SIDC is the number of clusters k is manually set and fixed for training. Setting an appropriate k is crucial to the model performance. Inspired by the DBSCAN (Ester et al, 1996), different clusters can be decided based on the sample distribution density in a non-parametric way. The cluster number k is an outcome of DBSCAN and no need to be set beforehand. Therefore, combining SIDC with the idea of density-based clustering can potentially overcome such limitation.

Bibliography

Ahmed K, Baig MH, Torresani L (2016) Network of experts for large-scale image categorization. In: Proceedings of the European Conference on Computer Vision (ECCV)

Aljalbout E, Golkov V, Siddiqui Y, Cremers D (2018) Clustering with deep learning: Taxonomy and new methods. In: arXiv preprint

Andrew G, Arora R, Bilmes JA, Livescu K (2013) Deep canonical correlation analysis. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Bai S, Bai X, Tian Q (2017) Scalable person re-identification on supervised smoothed manifold. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research (JMLR) 7(Nov):2399–2434

Berinsky AJ, Huber GA, Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. Political analysis 20(3):351–368

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of the International Conference on Computational Statistics (COMPSTAT)

Cai D, He X, Han J (2010a) Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering (TKDE) 23(6):902–913

Cai D, He X, Han J, Huang TS (2010b) Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(8):1548–1560

Cai S, Zuo W, Zhang L (2017) Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Carlucci FM, Porzi L, Caputo B, Ricci E, Bulò SR (2017) Autodial: Automatic domain alignment layers. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Chandar S, Khapra MM, Larochelle H, Ravindran B (2016) Correlational neural networks. Neural computation 28(2):257–285

Chang CC, Lin CJ (2011) Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):27

Chechik G, Sharma V, Shalit U, Bengio S (2010) Large scale online learning of image similarity through ranking. Journal of Machine Learning Research (JMLR) 11(Mar):1109–1135

Chen C, Xie W, Huang W, Rong Y, Ding X, Huang Y, Xu T, Huang J (2019) Progressive feature alignment for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Chen H, Wang Y, Wang G, Qiao Y (2018) Lstd: A low-shot transfer detector for object detection. In: Association for the Advancement of Artificial Intelligence (AAAI)

Cheung B, Livezey JA, Bansal AK, Olshausen BA (2015) Discovering hidden factors of variation in deep networks. In: Proceedings of the International Conference on Learning Representations Workshop (ICLRW)

Cogswell M, Ahmed F, Girshick R, Zitnick L, Batra D (2015) Reducing overfitting in deep networks by decorrelating representations. In: arXiv preprint

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Deng W, Zheng L, Kang G, Yang Y, Ye Q, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Du M, Sankaranarayanan AC, Chellappa R (2014) Robust face recognition from multi-view videos. IEEE Transactions on Image Processing (TIP) 23(3):1105–1117

Eigen D, Ranzato M, Sutskever I (2014) Learning factored representations in a deep mixture of experts. In: International Conference on Learning Representations Workshop (ICLRW)

Ester M, Kriegel HP, Sander J, Xu X, et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of Conference on Knowledge Discovery and Data Mining (KDD)

Fan H, Mei X, Prokhorov D, Ling H (2018a) Multi-level contextual rnns with attention model for scene labeling. IEEE Transactions on Intelligent Transportation Systems 19(11):3475–3485

Fan H, Zheng L, Yan C, Yang Y (2018b) Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 14(4):83

Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32(9):1627–1645

Ferrari V, Zisserman A (2008) Learning visual attributes. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Freeman WT, Roth M (1995) Orientation histograms for hand gesture recognition. In: International workshop on automatic face and gesture recognition

Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016) Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) 17(1):2096–2030

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)

Golub GH, Zha H (1995) The canonical correlations of matrix pairs and their numerical computation. In: Linear algebra for signal processing, Springer, pp 27–49

Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. International Journal of Computer Vision (IJCV) 106(2):210–233

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Gretton A, Fukumizu K, Harchaoui Z, Sriperumbudur BK (2009) A fast, consistent kernel two-sample test. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Gross R, Matthews I, Cohn J, Kanade T, Baker S (2007) The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Carnegie Mellon University Robotics Institute TR-07–08

Guo Y, Xiao M (2012) Cross language text classification via subspace co-regularized multi-view learning. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural computation 16(12):2639–2664

Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

He K, Zhang X, Ren S, Sun J (2016a) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

He K, Zhang X, Ren S, Sun J (2016b) Identity mappings in deep residual networks. In: Proceedings of the European Conference on Computer Vision (ECCV)

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

Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros AA, Darrell T (2017) Cycada:

Cycle-consistent adversarial domain adaptation. In: Proceedings of Machine Learning Research (PMLR)

Hofmann T, Schölkopf B, Smola AJ (2008) Kernel methods in machine learning. The annals of statistics pp 1171–1220

Hotelling H (1936) Relations between two sets of variates. Biometrika

Hou S, Liu X, Wang Z (2017) Dualnet: Learn complementary features for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Hsiao WL, Grauman K (2017) Learning the latent. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: arXiv preprint

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Computation 3(1):79–87

Jang E, Gu S, Poole B (2017) Categorical reparameterization with gumbel-softmax. In: Proceedings of the International Conference on Learning Representations (ICLR)

Jia K, Sun L, Gao S, Song Z, Shi BE (2015) Laplacian auto-encoders: An explicit learning of nonlinear data manifold. Neurocomputing 160:250–260

Jiang Z, Zheng Y, Tan H, Tang B, Zhou H (2017) Variational deep embedding: An unsupervised and generative approach to clustering. In: Proceedings of The International Joint Conference on Artificial Intelligence (IJCAI)

Jin X, Chen Y, Dong J, Feng J, Yan S (2016) Collaborative layer-wise discriminative learning in deep neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV)

Karaletsos T, Belongie S, Rätsch G (2015) Bayesian representation learning with oracle constraints. In: arXiv preprint

Khamis S, Kuo CH, Singh VK, Shet VD, Davis LS (2014) Joint learning for attribute-consistent person re-identification. In: Proceedings of the European Conference on Computer Vision Workshop (ECCVW)

Kim T, Cha M, Kim H, Lee JK, Kim J (2017) Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Kingma D, Ba J (2014) Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR)

Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: Proceedings of the International Conference on Learning Representations (ICLR)

Kiwiel KC (2001) Convergence and efficiency of subgradient methods for quasiconvex minimization. Mathematical programming 90(1):1–25

Kodirov E, Xiang T, Fu Z, Gong S (2016) Person re-identification by unsupervised ℓ_1 graph learning. In: Proceedings of the European Conference on Computer Vision (ECCV)

Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. Technical report, University of Toronto

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS), pp 1097–1105

Kumar N, Berg A, Belhumeur PN, Nayar S (2011) Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(10):1962–1977

Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Lampert CH, Nickisch H, Harmeling S (2013) Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36(3):453–465

Layne R, Hospedales TM, Gong S, Mary Q (2012) Person re-identification by attributes. In: Proceedings of the British Machine Vision Conference (BMVC)

LeCun Y, Bottou L, Bengio Y, Haffner P, et al (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee CY, Xie S, Gallagher PW, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)

Li D, Chen X, Zhang Z, Huang K (2017a) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Li M, Zhu X, Gong S (2018a) Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European Conference on Computer Vision (ECCV)

Li S, Shao M, Fu Y (2017b) Person re-identification by cross-view multi-level dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40(12):2963–2977

Li SZ, Yi D, Lei Z, Liao S (2013) The casia nir-vis 2.0 face database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Li W, Zhu X, Gong S (2017c) Person re-identification by deep joint learning of multi-loss classification. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)

Li Y, Wu FX, Ngom A (2016) A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics 19(2):325–340

Li Y, Yang M, Mark Zhang Z (2018b) A survey of multi-view representation learning. IEEE Transactions on Knowledge and Data Engineering (TKDE) PP:1–1

Lin K, Yang HF, Hsiao JH, Chen CS (2015a) Deep learning of binary hash codes for fast image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)

Lin S, Li H, Li CT, Kot AC (2018) Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In: Proceedings of the British Machine Vision Conference (BMVC)

Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. Pattern Recognition (PR)

Lin Z, Ding G, Hu M, Wang J (2015b) Semantics-preserving hashing for cross-view retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017) Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Long J, Shelhamer E, Darrell T (2015a) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Long M, Cao Y, Wang J, Jordan MI (2015b) Learning transferable features with deep adaptation networks. In: arXiv preprint

Long M, Zhu H, Wang J, Jordan MI (2016) Unsupervised domain adaptation with residual transfer networks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Lowe DG, et al (1999) Object recognition from local scale-invariant features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Luo Z, Zou Y, Hoffman J, Fei-Fei LF (2017) Label efficient learning of transferable representations across domains and tasks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Maddison CJ, Mnih A, Teh YW (2017) The concrete distribution: A continuous relaxation of discrete random variables. In: Proceedings of the International Conference on Learning Representations (ICLR)

Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. In: arXiv preprint

Mathieu MF, Zhao JJ, Zhao J, Ramesh A, Sprechmann P, LeCun Y (2016) Disentangling factors of variation in deep representation using adversarial training. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Matsukawa T, Suzuki E (2016) Person re-identification using cnn features learned from combination of attributes. In: Proceedings of the International Conference on Pattern Recognition (ICPR)

McLaughlin N, del Rincon JM, Miller PC (2017) Person reidentification using deep convnets with multitask learning. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 27(3):525–539

Mendenhall WM, Sincich TL (2016) Statistics for Engineering and the Sciences. Chapman and Hall/CRC

Nadler B, Srebro N, Zhou X (2009) Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Nagarajan T, Grauman K (2018) Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV)

Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY (2011) Reading digits in natural images with unsupervised feature learning. In: Proceedings of the Conference on Advances in Neural Information Processing Systems Workshop (NIPSW)

Nocedal J, Wright SJ (2006) Numerical Optimization. Springer

Oja E (1983) Subspace methods of pattern recognition, vol 6. Research Studies Press

Van den Oord A, Schrauwen B (2014) Factoring variations in natural images with deep gaussian mixture models. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering (TKDE) 22(10):1345–1359

Pan X, Luo P, Shi J, Tang X (2018) Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV)

Parkhi OM, Vedaldi A, Zisserman A, et al (2015) Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC)

Peng P, Xiang T, Wang Y, Pontil M, Gong S, Huang T, Tian Y (2016) Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision Workshop (ECCVW)

Romera-Paredes B, Torr P (2015) An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning (ICML)

Rozantsev A, Salzmann M, Fua P (2018) Residual parameter transfer for deep domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Saito K, Ushiku Y, Harada T (2017) Asymmetric tri-training for unsupervised domain adaptation. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Salakhutdinov R, Hinton G (2009) Semantic hashing. International Journal of Approximate Reasoning 50(7):969–978

Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) 35(4):119

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Schumann A, Stiefelhagen R (2017) Person re-identification by deep learning attribute-complementary information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

Shaham U, Stanton K, Li H, Nadler B, Basri R, Kluger Y (2018) Spectralnet: Spectral clustering using deep neural networks. In: Proceedings of the Conference on International Conference on Learning Representations (ICLR)

Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J (2017) Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Proceedings of the International Conference on Learning Representations (ICLR)

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the Conference on International Conference on Learning Representations (ICLR)

Sohn K, Liu S, Zhong G, Yu X, Yang MH, Chandraker M (2017) Unsupervised domain adaptation for face recognition in unlabeled videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research (JMLR) 15(1):1929–1958

Sun SH, Huh M, Liao YH, Zhang N, Lim JJ (2018) Multi-view to novel view: Synthesizing novel views with self-learned confidence. In: Proceedings of the European Conference on Computer Vision (ECCV)

Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015a) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–9

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015b) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Takagi M, Simo-Serra E, Iizuka S, Ishikawa H (2017) What makes a style: Experimental analysis of fashion prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Tao D, Guo Y, Yu B, Pang J, Yu Z (2017) Deep multi-view feature learning for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 28(10):2657–2666

Tian Y, Peng X, Zhao L, Zhang S, Metaxas DN (2018) Cr-gan: learning complete representations for multi-view generation. In: arXiv preprint

Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Uurtio V, Monteiro JM, Kandola J, Shawe-Taylor J, Fernandez-Reyes D, Rousu J (2018) A tutorial on canonical correlation methods. ACM Computing Surveys (CSUR) 50(6):95

Varior RR, Haloi M, Wang G (2016a) Gated siamese convolutional neural network architecture for human re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV)

Varior RR, Shuai B, Lu J, Xu D, Wang G (2016b) A siamese long short-term memory architecture for human re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV)

Veit A, Belongie S, Karaletsos T (2017) Conditional similarity networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Vinyals O, Blundell C, Lillicrap T, Wierstra D, et al (2016) Matching networks for one shot learning. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Viroli C, McLachlan GJ (2017) Deep gaussian mixture models. In: Statistics and Computing

Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Wang M, Deng W (2018) Deep face recognition: A survey. In: arXiv preprint

Wang W, Arora R, Livescu K, Bilmes J (2015a) On deep multi-view representation learning. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Wang W, Arora R, Livescu K, Bilmes JA (2015b) Unsupervised learning of acoustic features via deep canonical correlation analysis. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

Wang W, Arora R, Livescu K, Srebro N (2015c) Stochastic optimization for deep cca via nonlinear orthogonal iterations. In: Allerton

Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Wen Y, Li Z, Qiao Y (2016) Latent factor guided convolutional neural networks for age-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Xiao T, Li S, Wang B, Lin L, Wang X (2017) Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Xiao Wang RYBLJT Shaofei Zheng (2019) Pedestrian attribute recognition: A survey. In: arXiv preprint

Xie J, Girshick R, Farhadi A (2016a) Unsupervised deep embedding for clustering analysis. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Xie S, Tu Z (2015) Holistically-nested edge detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Xie S, Girshick R, Dollár P, Tu Z, He K (2016b) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Xu C, Tao D, Xu C (2013) A survey on multi-view learning. In: arXiv preprint

Yanbei C, Xiatian Z, Shaogang G (2017) Person re-identification by deep learning multi-scale representations. In: Proceedings of the International Conference on Computer Vision Workshop (ICCVW)

Yang B, Fu X, Sidiropoulos ND, Hong M (2017a) Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In: Proceedings of the Conference on International Conference on Machine Learning (ICML)

Yang J, Parikh D, Batra D (2016a) Joint unsupervised learning of deep representations and image clusters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Yang S, Ramanan D (2015) Multi-scale recognition with dag-cnns. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Yang S, Li L, Wang S, Zhang W, Huang Q (2017b) A graph regularized deep neural network for unsupervised image representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Yang X, Molchanov P, Kautz J (2016b) Multilayer and multimodal fusion of deep neural networks for video classification. In: Proceedings of the ACM International Conference on Multimedia

Yeung KY, Ruzzo WL (2001) Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural

networks? In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Yu HX, Wu A, Zheng WS (2017a) Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Yu J, Tao D, Rui Y, Cheng J (2013) Pairwise constraints based multiview features fusion for scene classification. Pattern Recognition (PR) 46(2):483–496

Yu Q, Liu F, Song YZ, Xiang T, Hospedales TM, Loy CC (2016) Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Yu R, Zhou Z, Bai S, Bai X (2017b) Divide and fuse: A re-ranking approach for person re-identification. In: Proceedings of the British Machine Vision Conference (BMVC)

Yu W, Yang K, Yao H, Sun X, Xu P (2017c) Exploiting the complementary strengths of multi-layer cnn features for image retrieval. Neural Computing 237:235–241

Yue Z, Meng D, He J, Zhang G (2017) Semi-supervised learning through adaptive laplacian graph trimming. Image and Vision Computing 60:38–47

Yuksel SE, Wilson JN, Gader PD (2012) Twenty years of mixture of experts. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 23(8):1177–1193

Zhang J, Li W, Ogunbona P, Xu D (2019) Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. ACM Computing Surveys (CSUR) 52(1):7

Zhang Y, Xiang T, Hospedales TM, Lu H (2018) Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017a) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Zhao J, Xie X, Xu X, Sun S (2017b) Multi-view learning overview: Recent progress and new challenges. Information Fusion 38:43–54

Zhao L, Li X, Wang J, Zhuang Y (2017c) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Zheng L, Yang Y, Hauptmann AG (2016a) Person re-identification: Past, present and future. In: arXiv preprint

Zheng Z, Zheng L, Yang Y (2016b) A discriminatively learned cnn embedding for person re-identification. In: arXiv preprint

Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV)

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)

Zhu H, Long M, Wang J, Cao Y (2016) Deep hashing network for efficient similarity retrieval. In: Association for the Advancement of Artificial Intelligence (AAAI)

Zhu XJ (2005) Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences

Zhu Z, Luo P, Wang X, Tang X (2014) Multi-view perceptron: a deep model for learning face identity and view representations. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)