Knowledge Transfer in Object Recognition

Xu Lan

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

24 October 2020

TO MY PARENTS

Knowledge Transfer in Object Recognition

Xu Lan

Abstract

Object recognition is a fundamental and long-standing problem in computer vision. Since the latest resurgence of deep learning, thousands of techniques have been proposed and brought to commercial products to facilitate people's daily life. Although remarkable achievements in object recognition have been witnessed, existing machine learning approaches remain far away from human vision system, especially in learning new concepts and Knowledge Transfer (KT) across scenarios. One main reason is that current learning approaches address isolated tasks by independently training predefined models, without considering any knowledge learned from previous tasks or models. In contrast, humans have an inherent ability to transfer the knowledge acquired from earlier tasks or people to new scenarios. Therefore, to scaling object recognition in realistic deployment, effective KT schemes are required.

This thesis studies several aspects of KT for scaling object recognition systems. Specifically, to facilitate the KT process, several mechanisms on fine-grained and coarse-grained object recognition tasks are analyzed and studied, including 1) cross-class KT on person re-identification (re-id); 2) cross-domain KT on person re-identification; 3) cross-model KT on image classification; 4) cross-task KT on image classification. In summary, four types of knowledge transfer schemes are discussed as follows:

Chapter 3 *Cross-class* KT in person re-identification, one of representative fine-grained object recognition tasks, is firstly investigated. The nature of person identity classes for person re-id are totally disjoint between training and testing (a zero-shot learning problem), resulting in the highly demand of cross-class KT. To solve that, existing person re-id approaches aim to derive a feature representation for pairwise similarity based matching and ranking, which is able to generalise to test. However, current person re-id methods assume the provision of accurately cropped person bounding boxes and each of them is in the same resolution, ignoring the impact of the background noise and variant scale of images to cross-class KT. This is more severed in practice when person bounding boxes must be *detected automatically* given a very large number of images and/or videos (un-constrained scene images) processed. To address these challenges, this chapter provides two novel approaches, aiming to promote cross-class KT and boost re-id performance. 1) This chapter alleviates inaccurate person bounding box by developing a joint learning deep model that optimises person re-id attention selection within any auto-detected person bounding boxes by *reinforcement learning* of background clutter minimisation. Specifically, this chapter formulates a novel unified re-id architecture called **I**dentity **D**iscriminativ**E** **A**ttention reinforcement **L**earning (IDEAL) to accurately select re-id attention in auto-detected bounding boxes for optimising re-id performance. 2) This chapter addresses multi-scale problem by proposing a *Cross-Level Semantic Alignment* (CLSA) deep learning approach capable of learning more discriminative identity feature representations in a unified end-to-end model. This

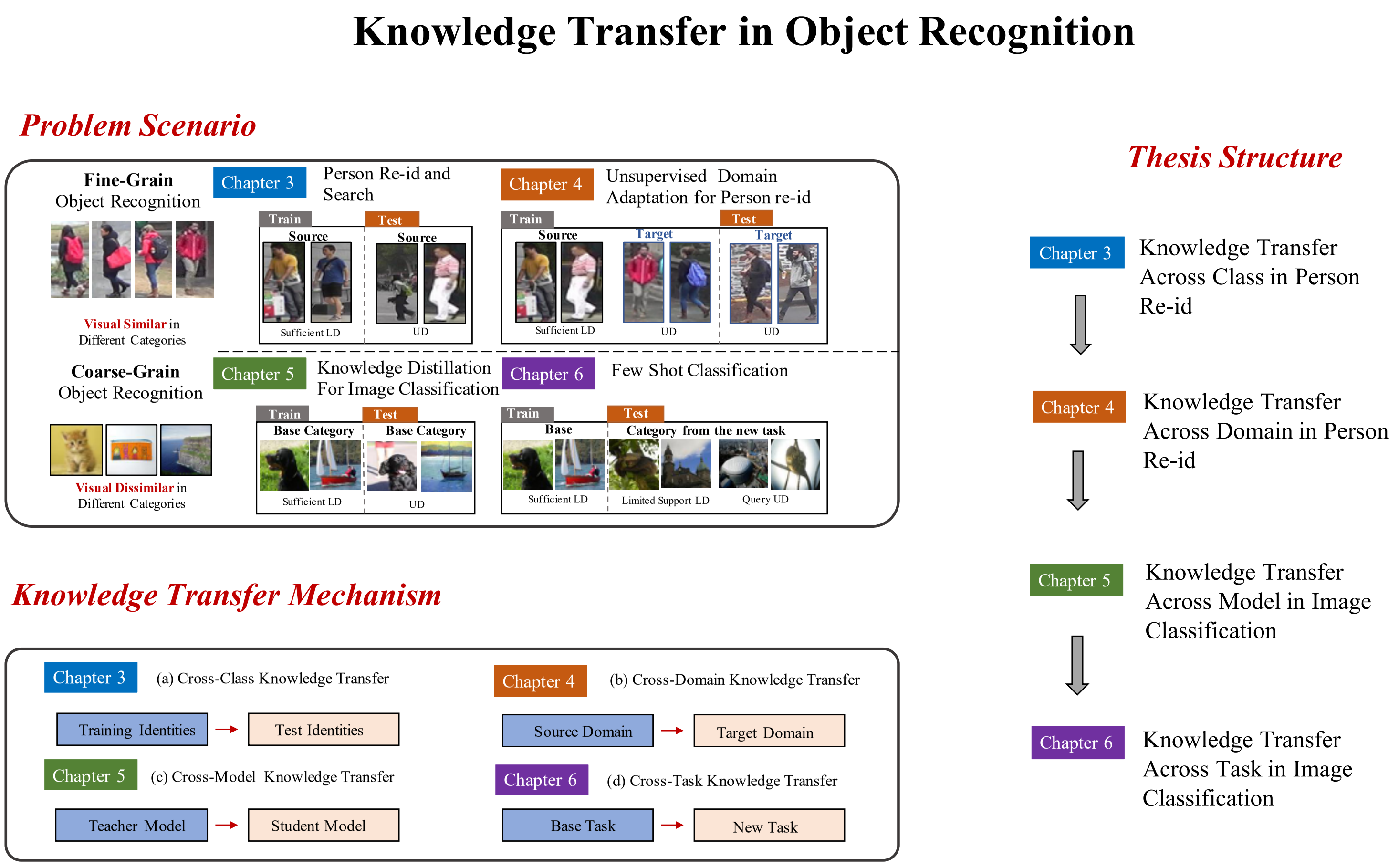
is realised by exploiting the in-network feature pyramid structure of a deep neural network enhanced by a novel cross pyramid-level semantic alignment loss function. Extensive experiments show the modelling advantages and performance superiority of both IDEAL and CLSA over the state-of-the-art re-id methods on widely used benchmarking datasets.

Chapter 4 In this chapter, we address the problem of *cross-domain* KT in unsupervised domain adaptation for person re-id. Specifically, this chapter considers cross-domain KT as follows: 1) Unsupervised domain adaptation: "*train once, run once*" pattern, transferring knowledge from source domain to specific target domain and the model is restricted to be applied on target domain only; 2) Universal re-id: "*train once, run everywhere*" pattern, transferring knowledge from source domain to any target domains, and therefore is capable of deploying any domains of re-id task. This chapter firstly develops a novel Hierarchical Unsupervised Domain Adaptation (HUDA) method for unsupervised domain adaptation for re-id. It can automatically transfer labelled information of an existing dataset (a source domain) to an unlabelled target domain for unsupervised person re-id. Specifically, HUDA is designed to model jointly global distribution alignment and local instance alignment in a two-level hierarchy for discovering transferable source knowledge in unsupervised domain adaptation. Crucially, this approach aims to overcome the under-constrained learning problem of existing unsupervised domain adaptation methods, lacking of the local instance alignment constraint. The consequence is more effective and cross-domain KT from the labelled source domain to the unlabelled target domain. This chapter further addresses the limitation of "*train once, run once* " for existing domain adaptation person re-id approaches by presenting a novel "*train once, run everywhere*" pattern. This conventional "train once, run once" pattern is unscalable to a large number of target domains typically encountered in real-world deployments, due to the requirement of training a separate model for each target domain as supervised learning methods. To mitigate this weakness, a novel "*Universal Model Learning*" (UML) approach is formulated to enable domain-generic person re-id using only limited training data of a "*single*" seed domain. Specifically, UML trains a universal re-id model to discriminate between a set of transformed person identity classes. Each of such classes is formed by applying a variety of random appearance transformations to the images of that class, where the transformations simulate camera viewing conditions of any domains for making the model domain generic.

Chapter 5 The third problem considered in this thesis is *cross-model* KT in coarse-grained object recognition. This chapter discusses knowledge distillation in image classification. Knowledge distillation is an effective approach to transfer knowledge from a large teacher neural network to a small student (target) network for satisfying the low-memory and fast running requirements. Whilst being able to create stronger target networks compared to the vanilla non-teacher based learning strategy, this scheme needs to train additionally a large teacher model with expensive computational cost and requires complex multi-stages training. This chapter firstly presents a Self-Referenced Deep Learning (SRDL) strategy to accelerate the training process. Unlike both vanilla optimisation and knowledge distillation, SRDL distils the knowledge discovered by the in-training target model back to itself for regularising the subsequent learning procedure therefore eliminating the need for training a large teacher model. Secondly, an On-the-fly Native Ensemble (ONE) learning strategy for one-stage knowledge distillation is proposed to solve the weakness of complex multi-stages training. Specifically, ONE only trains a single multi-branch network while simultaneously establishing a strong teacher on-the-fly to enhance the learning of target network.

Chapter 6 Forth, this thesis studies the *cross-task* KT in coarse-grained object recognition. This chapter focuses on the few-shot classification problem, which aims to train models capable of recognising new, previously unseen categories from the novel task by using only limited training samples. Existing metric learning approaches constitute a highly popular strategy, learning discriminative representations such that images, containing different classes, are well separated in an embedding space. The commonly held assumption that each class is summarised by a single, global representation (referred to as a prototype) that is then used as a reference to infer class labels brings significant drawbacks. This formulation fails to capture the complex multi-modal latent distributions that often exist in real-world problems, and yields models that are highly sensitive to the prototype quality. To address these limitations, this chapter proposes a novel Mixture of Prototypes (MP) approach that learns multi-modal class representations, and can be integrated into existing metric based methods. MP models class prototypes as a group of feature representations carefully designed to be highly diverse and maximise ensembling performance. Furthermore, this thesis investigates the benefit of incorporating unlabelled data in cross-task KT, and focuses on the problem of Semi-Supervised Few-shot Learning (SS-FSL). Recent SS-FSL work has relied on popular Semi-Supervised Learning (SSL) concepts, involving iterative pseudo-labelling, yet often yields models that are susceptible to error propagation and sensitive to initialisation. To address this limitation, this chapter introduces a novel prototype-based approach (Fewmatch) for SS-FSL that exploits model Consistency Regularization (CR) in a robust manner and promotes cross-task unlabelled data knowledge transfer. Fewmatch exploits unlabelled data via Dynamic Prototype Refinement (DPR) approach, where novel class prototypes are alternatively refined 1) explicitly, using unlabelled data with high confidence class predictions and 2) implicitly, by model fine-tuning using a data selective model CR loss. DPR affords CR convergence, with the explicit refinement providing an increasingly stronger initialisation and alleviates the issue of error propagation, due to the application of CR.

Chapter 7 draws conclusions and suggests future works that extend the ideas and methods developed in this thesis.


Knowledge Transfer in Object Recognition
Problem Scenario
Fine-Grain
Object Recognition
Visual Similar in
Different Categories
Chapter 3
Person Re-id and
Search
Train
Test
Source
Source
Sufficient LD
UD
Chapter 4
Unsupervised Domain
Adaptation for Person re-id
Train
Test
Source
Target
Target
Sufficient LD
UD
UD
Coarse-Grain
Object Recognition
Visual Dissimilar in
Different Categories
Chapter 5
Knowledge Distillation
For Image Classification
Train
Test
Base Category
Base Category
Sufficient LD
UD
Chapter 6
Few Shot Classification
Train
Test
Base
Category from the new task
Sufficient LD
Limited Support LD
Query UD
Knowledge Transfer Mechanism
Chapter 3
(a) Cross-Class Knowledge Transfer
Training Identities
Test Identities
Chapter 4
(b) Cross-Domain Knowledge Transfer
Source Domain
Target Domain
Chapter 5
(c) Cross-Model Knowledge Transfer
Teacher Model
Student Model
Chapter 6
(d) Cross-Task Knowledge Transfer
Base Task
New Task
Thesis Structure
Chapter 3
Knowledge Transfer
Across Class in Person
Re-id
Chapter 4
Knowledge Transfer
Across Domain in Person
Re-id
Chapter 5
Knowledge Transfer
Across Model in Image
Classification
Chapter 6
Knowledge Transfer
Across Task in Image
Classification

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 parts of the work have previously been published or in submission as:

Chapter 3

1. **Xu Lan**, HanXiao Wang, Shaogang Gong and Xiatian Zhu. Deep Reinforcement Learning Attention Selection for Person Re-identification. The 28th British Machine Vision Conference (**BMVC2017**), London, UK, September 2017.

2. **Xu Lan**, Xiatian Zhu, Shaogang Gong. Person Search by Multi-Scale Matching. The 15th European Conference on Computer Vision (**ECCV2018**), Munich, Germany, September 2018.

Chapter 4

1. **Xu Lan**, Xiatian Zhu, Shaogang Gong. Unsupervised Cross-Domain Person Re-Identification by Instance and Distribution Alignment. IEEE Transactions on Image Processing. (**TIP in submission**

2. **Xu Lan**, Xiatian Zhu, Shaogang Gong. Univeral Person re-identification. Pattern Recognition. (**PR**, in submission)

Chapter 5

1. **Xu Lan**, Xiatian Zhu, Shaogang Gong. Self-Referenced Deep Learning, The 14th Asian Conference on Computer Vision (**ACCV2018**), Perth, Australia, December 2018.

2. **Xu Lan**, Xiatian Zhu, Shaogang Gong. Knowledge Distillation by On-the-Fly Native Ensemble, The 32th Conference on Neural Information Processing Systems (**NeurIPS2018**), Montréal, Canada, December 2018.

Chapter 6

1. **Xu Lan**, Steven McDonagh, Shaogang Gong, Aoxue Li, Weiran Huang, Jiashi Feng, Sarah Parisot. Learning Diverse and Representative Proxy Mixtures for Few-shot Classification.

35th AAAI Conference on Artificial Intelligence (**AAAI 2021, in submission**).

2. **Xu Lan**, Steven McDonagh, Shaogang Gong, Jiali Wang , Zhenguo Li, Sarah Parisot. FewMatch: Prototype based Consistency Regularization for Semi-Supervised Few-Shot Classification. The 34th Conference on Neural Information Processing Systems (**NeurIPS2020, in submission**).

Acknowledgements

First of all, I would like to express my greatest gratitude to my main supervisor Prof. Shaogang Gong for his perpetual patience, continued encouragement and support in the last four years. I appreciate it so much he gives me the freedom to chose the topic I prefer and is willing to provide any help once it is benefical to my research and daily life. The brainstorm in front of the office's whiteboard and the time of we burn the midnight oil to catch up the deadline, are fond memories of my life. Meanwhile, I convey my special thanks to Dr. Xiatian Zhu for his excellent guidance and invaluable advice on my Ph.D. projects and thesis writing. I will never forget his encouragement when our paper is rejected. This helps me a lot to go through the dark time and inspires me to keep climbing and never give up. I also wish to thank my academic progression panel members Tao Xiang and Ioannis Patras for their constructive and professional suggestions on my research direction.

My warm appreciation goes to all members I met at Vision Group in QMUL for their kindness and concern: Hanxiao Wang, Wei Li, JingYa Wang, Qi Dong, Hang Su, Zhiyi Cheng, Yanbei Chen, Jiabo Huang, Guile Wu, Qian Yu, Li Zhang, Jifei Song, Kaiyue Peng, Conghui Hu, Ying Zhang. I am very grateful to the EECS administrative staff and system supports staff, especially Mellisa Yao, Tim Kay and Harry Krikelis to take care of my Ph.D. progress and provide the unbelievable 24/7 GPU server support. Besides, I shall thank the researchers outside QMUL: Sarah Parisot, Steven McDonagh for their incredible advise and help for my Ph.D. project.

My deep and sincere gratitude to my parents for continuous and unconditional love. It is ten years now since I am away from my hometown to pursue my study. No matter how far I away from them, they are always my backbone. At last, I would like to give special thanks to my loving fiancée Jiali Wang for her endless devotion and support without any complaints. I will never forget her sacrifice of leaving the comfortable working environment and giving up valuable job in china, in order to accompany with me in london.

Contents

List of Figures

List of Acronyms

KT	Knowledge Transfer
IDEAL	Identity DiscriminativE Attention reinforcement Learning
CLSA	Cross-Level Semantic Alignment
HUDA	Hierarchical Unsupervised Domain Adaptation
UML	Universal Model Learning
AGI	Artificial General Intelligence
SRDL	Self-Referenced Deep Learning
ONE	On-the-fly Native Ensemble
MP	Mixture of Prototypes
FSL	Few Shot Learning
FSC	Few Shot Classification
CNN	Convolutional Neural Network
NAS	Neural Architecture Search
SSFSC	Semi-Supervised Few Shot Classification
CR	Consistency Regularization

Chapter 1

Introduction

The only source of knowledge is experience

—— **Albert Einstein**

"Can machines think?" This is one of the famous questions asked by Alan Turing in 1950 and inspired generations of researchers to set the foundations for Artificial Intelligence (AI) and other related sub-fields in computer science. Over the last few decades, it is endowed more specific and realistic meaning that whether the machine system has the capability of understanding or addressing any intellectual tasks as a human being can [1], which is known as Artificial General Intelligence (AGI). To fulfill AGI, there are several traits that an intelligent system should have, including common sense, knowledge transfer (transfer learning), conceptual understanding, causality, etc. Among them, Demis Hassabis points out that the key to AGI lies in knowledge transfer.

However, existing AI systems often independently solved these problems in isolation, without exploring systems full generality and utilizing the previous acquired knowledge. As a result, such system design becomes excessively brittle and usually fails when an unexpected circumstance appears. In contrast, when human beings encounter new scenarios, they firstly search their memories for anything which matches a set of conditions and transfer the corresponding knowledge to the current scenario.

To mirror the way of humans addressing the intellectual tasks, this thesis focuses on the problem of knowledge transfer in computer vision. In particular, this thesis studies knowledge transfer

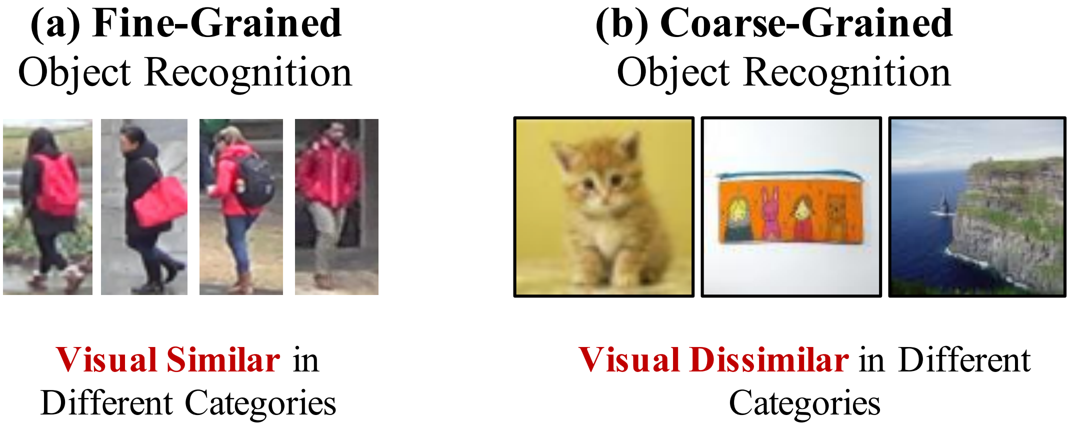


Figure 1.1: Illustration of object recognition. (a) Fine-grained object recognition, each of categories is visual similar. (b) Coarse-grained object recognition, each of categories is visual dissimilar.

in object recognition, which is the primary and crucial task for computer vision. By means of effective knowledge transfer, we have the capacity to scale the existing object recognition systems towards more realistic deployments.

1.1 Object Recognition

Object recognition is a general term to describe a collection of related computer vision tasks that involve identifying objects in digital photographs [2]. It roughly contains both image classification (a task requiring an algorithm to determine what object classes are present in the image) as well as object detection (a task requiring an algorithm to localize all objects present in the image) [3]. Object recognition is a fundamental task in computer vision and has been an active research area for several decades. According to the visual similarity of categories, the object recognition could be summarized in two types: 1) coarse-grained object recognition 2) fine-grained object recognition (Figure 1.1). The former deals with more general category recognition, which category is the basic level, such as dog, car and house, while the latter aims to distinguish subordinate-level categories within a general category, such as different species of bird or different human identities. This task is extremely challenging due to high intra-class and low inter-class variance.

Designing a scalable visual recognition system in more complex and realistic situations,

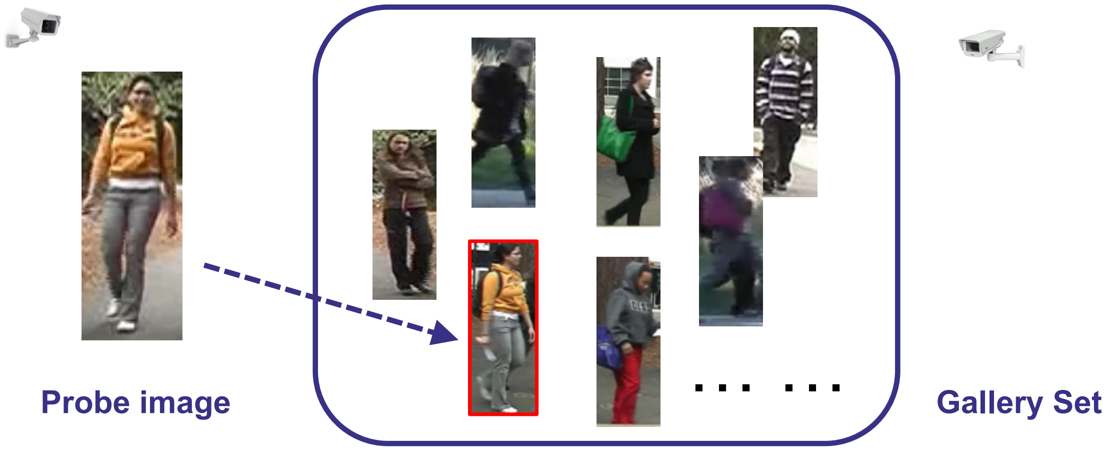


Figure 1.2: Illustration of person re-identification: recognizing an individual n diverse locations, different times and over different non-overlapping camera views.

which surpasses human, is both academic and industry community desperate for. Despite deep learning models, in particular, in the context of image classification, can reach super-human performance on some benchmarks when trained on the large amounts of annotated data [4, 5], these models still have a long way to matching humans behaviour when unexpected and novel scenarios appear, due to lacking of the human knowledge transfer mechanism.

In this thesis, we discuss two representative sub-fields in object recognition: fine-grained person re-identification and coarse-grained image classification, with the goal of scaling the existing visual recognition system in more realistic scenarios.

1.1.1 Person Re-identification

Person re-identification (re-id) is one of the representative fine-grained object recognition tasks, aiming at searching people across non-overlapping camera views distributed at different locations (gallery set) by matching person bounding box images in probe set [6] (Figure 1.2). Re-id has become a fundamental technique for existing intelligent surveillance systems and plays a critical role in ensuring public security and assisting people daily life. For example, the government deploys re-id systems in closed-circuit television (CCTV) camera networks to fast locate suspicious criminals, therefore protecting public order and civilian safety.

Although impressive achievements have been obtained by deep learning in visual recognition, person re-id still remains a non-trivial and challenging task. This is due to the person identities non-overlap between training and testing, requiring the high demand of cross-class knowledge

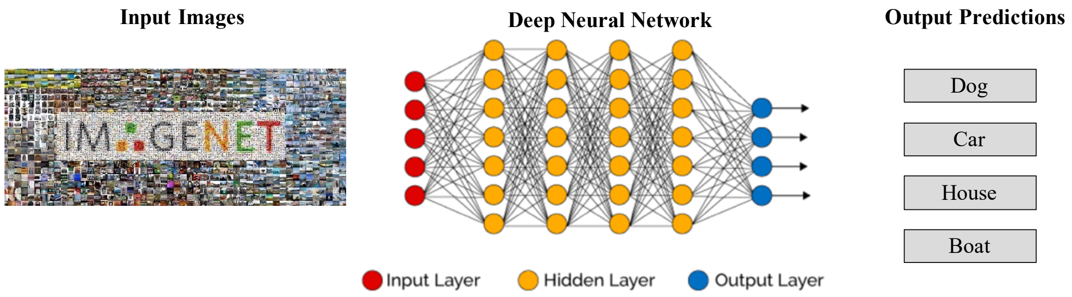


Figure 1.3: Illustration of image classification by deep neural networks on ImageNet. [3].

transfer. However, the transferring process is very brittle to the several complicated factors such as the variant scale of images (distances between the camera and the individual is uncontrolled), noisy background, etc. Section 1.3.1 and 1.3.2 detail these challenges and provide effective solutions to alleviate these weaknesses.

1.1.2 Image Classification

Image classification is a very challenging task, mainly due to a large amount of intra-class variability, arising from different lightings, occlusions, background and corruptions, etc. Recently, deep convolutional neural networks have led to a series of breakthroughs for image classification [5, 7, 8]. The core idea of deep learning is to discover multiple levels of representation (Figure 1.3), with the hope that higher-level features encode abstract semantic information of the data, which is expected to provide more invariance to the intra-class variability. To learn a powerful representational feature, various deep nerual architectures are designed, including VGG [4], Inception [9], ResNet [5], and DenseNet [10], etc.

While large capacity models often achieve competitive results, the expensive time cost, especially in real-time deployment, leads to obvious shortcomings in realistic. Besides, the well-trained model still suffers from the intrinsic drawback of the machine learning model without considering knowledge transfer, resulting in the limited generalization in new scenarios. In this thesis, Section 1.3.3 and 1.3.4 detail these weaknesses and provide alternative solutions.

1.2 Knowledge Transfer

Knowledge Transfer (KT) refers to the learning paradigm in which an algorithm extracts knowledge from one or more application scenarios to help boost the learning performance in the other scenario [11]. One of the basic transfer learning techniques in the deep learning context is fine-tuning. Fine-tuning uses the pre-trained model which is trained on the large dataset *i.e*. ImageNet, and re-trains it on the target task with a very low learning rate. However, recently Kaiming *et al*. [12] argue pre-training such on ImageNet only speeds up convergence early in training, while does not necessarily boost final target task accuracy. This indicates that simply fine-tuning is not an effective KT strategy. The ineffectiveness of KT has obvious disadvantages including: 1) The model is short of the generalization ability when a new scenario appears, yielding the model often fails to recognize the novel object, due to ignoring the knowledge acquired from previous learning processes. To address this issue, cross-"**X**" knowledge transfer is required. Given different contexts, X could be classes, domains or tasks, etc. 2) The extensibility of the model is very limited. Supposing a new model is designed, we need to train it independently from scratch, while the knowledge learned from the existing well-trained model is ignored. In many real-life applications, deploying a very large capacity deep model is unrealistic, since these models are often computationally expensive with millions of parameters [13, 5, 4]. Therefore, it might not satisfy the real-time deployment requirement. However, training the low capacity model independently is insufficient, as it lacks ability to capture the complex data distribution in the real world scenarios, compared to large capacity models. In the last few years, researchers start to investigate whether it is possible to transfer knowledge between different models to emulate the human education systems, *i.e*. the student learns the knowledge from teachers. In summary, this problem could be concluded in one of the special cases for cross-"**X**" knowledge transfer, where "**X**" represents the deep learning model.

Recently, many researchers implicitly or explicitly point out the aforementioned issues. The related technique or problem to cross-"**X**" knowledge transfer could be summarized as: 1) cross-class knowledge transfer in person re-id; 2) cross-domain knowledge transfer in domain adaptation, 3) cross-model knowledge transfer in knowledge distillation, 4) cross-task knowledge transfer in few shot classification. As the cross-class knowledge transfer is discussed in Section 1.1.1, the rest of cross-"**X**" knowledge is provided as follows.

1.2.1 Domain Adaptation

Domain Adaptation (DA) is an area related to the cross-domain knowledge transfer, in which there is a domain shift or a distribution change between the source and target domain [15] (See Figure 1.4). While most of the existing visual recognition systems assume the training and eval-

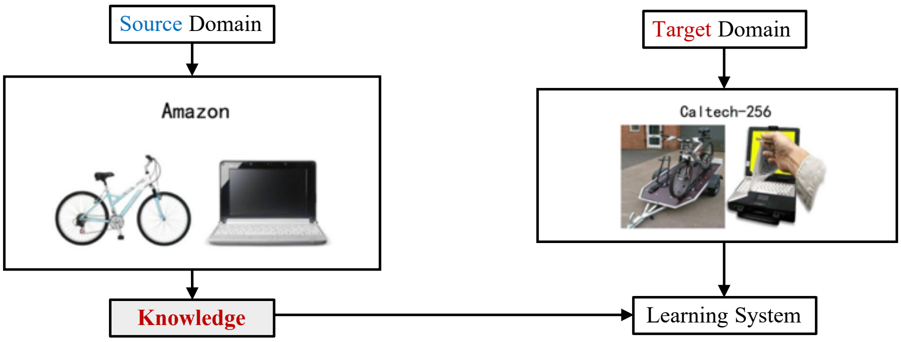


Figure 1.4: Some image examples from the ”Bike” and ”Laptop” categories in Amazon, and Caltech-256 databases [14]. There is a data distribution change between Amazon and Caltech-256 dataset. Cross-domain knowledge transfer is required. If the model is trained on the source domain (*i.e*. Amazon) and aims to deploy on the target domain (*i.e*. Caltech-256).

uation share the same data distribution, real-world deployment often encounters the distribution changes, known as the domain shift. In general, the domain shift might come from a set of complex factors, including background clutter, intra-category variation, motion blur, and scene illumination, etc. For example, in person re-id, data collected from Asian people is biased with respect to the total human population. If the data is biased, then the re-id system will be likely to make more mistakes to identify the people who are not from Asia. With the breakthrough on a various types of visual recognition tasks offered by deep learning, researchers find the deep neural networks have the capacity to learn transferrable features across tasks, alleviating the domain shift in DA. The basic idea is to add some regularization when source feature learning, aiming to well generalize to the new domain. This challenge is further discussed in Section 1.3.2.

1.2.2 Knowledge Distillation

Deep neural networks have gained substantial success in many computer vision tasks [16, 4, 17, 5, 18, 19, 20, 21]. However, the performance advantages often come at the cost of training and deploying resource-intensive networks with large depth and/or width [13, 5, 4]. This has given rise to efforts in developing more compact models, such as parameter binarisation [22], filter pruning [23], model compression [24], and knowledge distillation [25]. Among these existing techniques, knowledge distillation [25] is a generic cross-model knowledge transfer approach suitable to a wide variety of networks and applications. Knowledge distillation is an effective approach to transferring knowledge from a teacher neural network to a student target network for

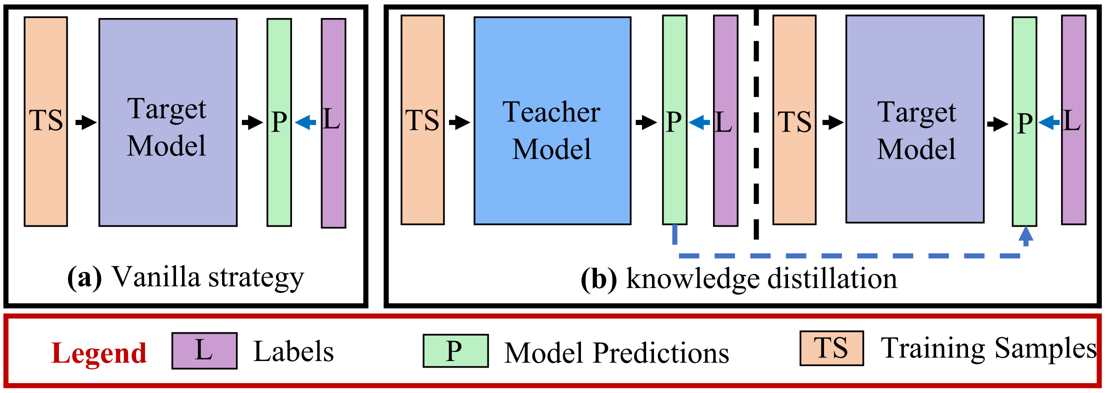


Figure 1.5: Illustration of two different deep network learning methods. **(a)** The vanilla training: Optimise the target model from the supervision of training label for M epochs in one stage. **(b)** The Knowledge Distillation training: Firstly learn a teacher model (the same architecture as the target model, but it is well-trained in contrast to the target model) in a *computationally intensive* manner; Then extract the learned knowledge from the teacher model; Lastly optimise the target model by leveraging both the label data and the teacher's knowledge for M epochs.

satisfying the low-memory and fast running requirements in practice use.

As a solution to this challenge, knowledge distillation first trains a deeper and/or wider "teacher" network (or an ensemble model), then learns a smaller "student" network to imitate the teacher's classification probabilities [25] and/or feature representations [26, 27] (Figure 1.5 (b)). This imposes additional information beyond conventional supervised learning signals (Figure 1.5 (a)), leading to a more discriminative student model than learning the target model without the teacher's knowledge. Whilst being able to create stronger target networks compared to the vanilla non-teacher based learning strategy, this scheme needs to train *additionally* a large teacher model with the expensive computational cost. To that end, this thesis details this challenge in Section 1.3.3.

1.2.3 Few Shot Learning

Deep neural network beats humans at classifying images from ImageNet [5]. At that point, one could argue that computers become better than human when there is a large amount of annotated data. However, this is not realistic to collect such large scale of the dataset, *i.e.* ImageNet, for every task. Sometimes, there are only one or two labelled examples per category. In this context, the deep model often experiences persistent failures and frustrations. This scenario is

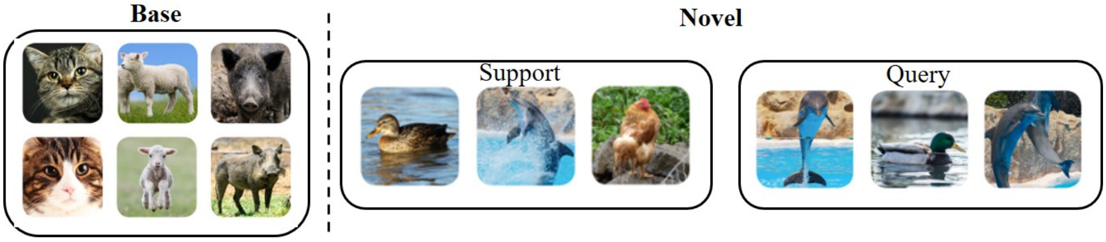


Figure 1.6: The illustration of FSL classification problem. The base category contains sufficient labelled samples, while the model aims at generalizing well on novel category from query set in new tasks with limited labelled samples available in the support set.

known as Few-Shot Learning (FSL). FSL aims to emulate human behaviour by teaching models to recognise and handle unseen classes from the new task in data-limited regimes. This requires highly cross-task knowledge transfer to enable the model to perform well on the new task.

For a few years now, the few-shot learning problem has drawn a lot of attention in the research community. In the FSL, researchers often access a large scale training dataset with sufficient annotated samples comprised base categories. The aim of FSL is obtaining a model which fast adapts to the new task. Each of novel categories from the new task is associated with only a few K labelled samples (e.g. ≤ 5 samples) compose the support set, while the remaining unlabelled samples consist the query set are used for evaluation (see Figure 1.6). Despite existing FSL approaches can boost cross-task knowledge transfer, there are still some common issues remained and Section 1.3.4 discusses these drawbacks.

1.3 Challenges, Solutions and Assumptions

1.3.1 Cross-class Knowledge Transfer on Person Re-id

Person re-id is the fine-grained object recognition task and each of categories is very similar, thus capturing discriminative features among visual similar classes is very essential. More importantly, the nature of disjoint person identities between training and testing requires highly desirable cross-class knowledge transfer, aiming to learn generalizable feature representation and boost re-id performance. To facilitate the cross-class knowledge transfer and mitigate negative influence of irrelevant factors, this thesis investigates two aspects in the context of person re-id in Chapter 3: the background and multi-scale of images.

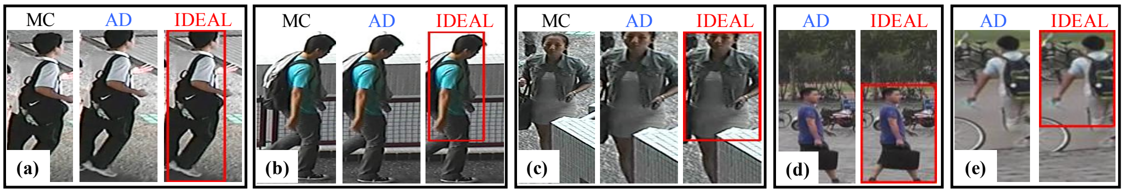


Figure 1.7: Comparisons of person bounding boxes by manually cropping (MC), automatically detecting (AD), and identity discriminative attention reinforcement learning (IDEAL). Often AD contains more background clutter (a,d,e). Both AD and MC may suffer from occlusion (c), or a lack of identity discriminative attention selection (b).

- *The background noise of person image box:* Existing person re-id methods assume the provision of accurately cropped bounding boxes (constrained scene) and little background noise, e.g. by *manually cropping*. This is inconsistent with real-world application scenarios where person bounding boxes can only be extracted less accurately by *automatic person detection*, given limited human labelling budget and vast volume of surveillance video/image data. This inaccurately cropped bounding boxes, including noisy, increase the difficulty to transfer knowledge from train to test, especially the identity of training and testing is not overlapped. Therefore, post-detection bounding box refinement becomes inevitable for optimising re-id matching, which however is ignored in the literature.

- *Variant resolution of person bounding box:* In parallel to post-detection bounding box refinement, some researchers further propose to contain person detection with person re-id in a unified framework, resulting in a more realistic setting: person search. Person search aims to find a probe person in a gallery of whole unconstrained scene images [28] (See Figure 1.8 (a)). It is an extended form of person re-identification (re-id) [6] by additionally considering the requirement of automatically detecting people in the scene images besides matching the identity classes. Unlike the conventional person re-id problem assuming the gallery images as either manually cropped or carefully filtered auto-detected bounding boxes [29, 20, 30, 31, 32, 33, 34, 35, 36, 33], person search deals with raw unrefined detections with many false cropping and unknown degrees of misalignment. This yields a more challenging matching problem especially in the process of person re-id. Moreover, auto-detected person boxes often vary more significantly in scale (resolution) than the conventional person re-id benchmarks (Figure 1.8(b)), due to the inherent uncontrolled distances between persons and cameras (Figure 1.8(a)). To facilitate the feature

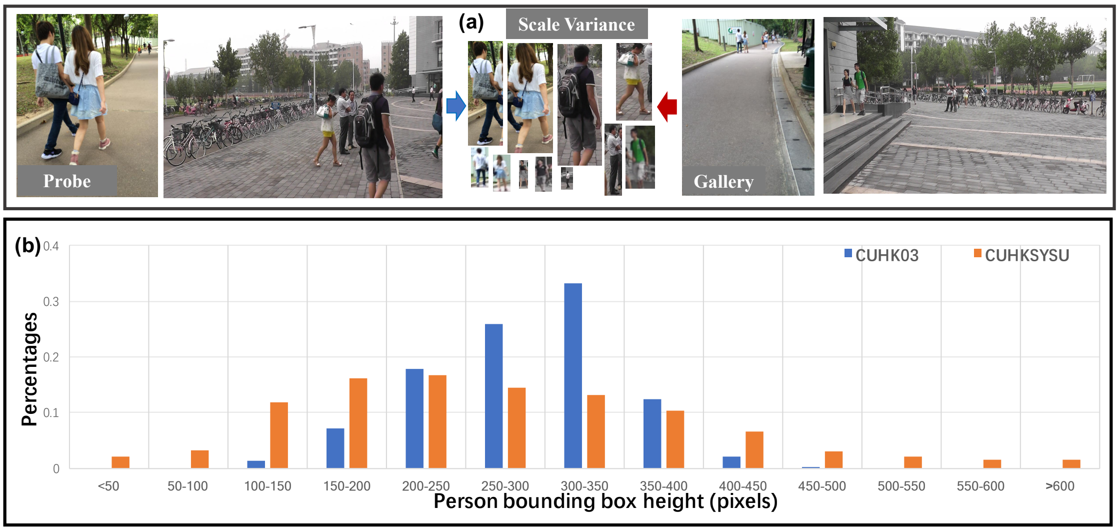


Figure 1.8: Illustration of the intrinsic multi-scale matching challenge in person search. (a) Auto-detected person bounding boxes vary significantly in scale. (b) The person scale distribution of CUHK-SYSU (person search benchmark) covers a much wider range than manually refined CUHK-03 (person re-id benchmark).

(knowledge) learned from training data successfully generalizing to test person matching, it is therefore intrinsic to consider *multi-scale feature learning* for cross-class knowledge transfer. However, this problem is currently under-studied in person search [28, 37, 38].

Solutions: This thesis provides solutions to address these two weaknesses. **(1)** To solve inaccurately cropped bounding boxes challenge, this thesis uses deep reinforcement learning which is a powerful technique that has been used in a wide range of application. In this work, we show the effectiveness of optimising auto-detected person bounding boxes in a deep reinforcement learning framework by aiming to automatically attend only re-id discriminative regions against complex background clutter. Specifically, a novel architecture called ***Identity DiscriminativE Attention reinforcement Learning*** (IDEAL) is formulated for accurate attention discovery in the context of auto-detected bounding boxes, capable of achieving similar re-id performance as compared to exhaustive manually labelling. **(2)** To alleviate the multi-scale matching problem, a ***Cross-Level Semantic Alignment*** (CLSA) deep learning approach is proposed to addressing the multi-scale matching challenge. This is based on learning an end-to-end in-network feature pyramid representation with superior robustness in coping with variable scales of auto-detected person bounding boxes.

Assumptions. The above solution is based on the following assumptions:

1. Auto-detected bounding boxes are not optimised for re-id tasks due to potentially more background clutter, occlusion, missing body part, and inaccurate bounding box alignment.

2. Attention selection is the need within poorer auto-detected bounding boxes as an integral part of learning to optimise person re-id accuracy in a fully automated process.

3. A single-scale feature representation is unable to capture the discriminative information at different scales which is useful to cross-class knowledge transfer in person identity matching;

4. A pyramid representation allows to be "scale-invariant" (or "scale insensitive") in the sense that a scale change in matching images is counteracted by a scale shift within the feature pyramid.

1.3.2 Cross-domain Knowledge Transfer on Person Re-id

This thesis studies cross-domain knowledge transfer in unsupervised domain adaptation for person re-id in Chapter 4. In particular, two weaknesses of existing unsupervised domain adaptation are addressed in this thesis.

- *Cross-domain knowledge transfer is insufficient due to lacking local instance alignment:* Most existing re-id methods rely heavily on *supervised learning* [34, 39, 40, 29, 41, 42, 43, 44], assuming that the model training and test data are drawn from the same camera network, i.e. the same domain. However, such trained models suffer from significant performance degradation when deployed to the unseen target domain due to the domain shift problem [45]. In reality, we often have *no* access to a large number of *manually* labelled matching person image pairs for every camera pair as required by supervised learning methods. Such large human labelling is both costly and not always available, due to a large number of camera pairs in each surveillance domain. Existing supervised learning methods have limited cross-domain usability. To overcome this limitation, a number of approaches have been proposed, including (1) hand-crafting features [46, 47], (2) image adaptation (synthesis) [48, 49, 50, 51], (3) feature adaptation [52, 53, 54, 55], (4) unsupervised deep learning [56], and (5) a hybrid of feature adaptation and unsupervised learning [50, 57, 58]. This thesis focuses on the *feature adaptation* approach for unsupervised cross-domain person re-id. The key idea is to align *feature statistics* between source and target training data. By doing so, re-id discriminative knowledge from the labelled source data can be transferred into the unlabelled target data. Existing feature adaptation methods typically rely on cross-domain alignment of *global feature distributions* [53, 54]. This however suffers from an *under-constrained optimisation* problem, which lacks local instance alignment constraints, yielding sub-optimal re-id models.

- *Cross-domain knowledge transfer required extra target domain training samples and well-trained models are only for the single domain deployment:* Whilst significant performance gains have been achieved on unlabelled target domains, existing unsupervised domain adaptation (Figure 1.9(b)), unsupervised model learning (Figure 1.9(c)), or their combination, often take a "*train once, run once*" pattern. That is, a trained model by them is effective only for the target domain that the model training is applied to. For every single target domain deployment, a new model needs to be trained through the same optimisation process repeatedly. Such a *domain-specific* property reduces their practical value and limits their scalability significantly, considering potentially a very large quantity of different domains to be targeted in real-world applications. Besides, the extra unlabelled target data is required to enable the adaptation. This results in extra cost for collecting these unlabelled data.

Solutions: To solve the aforementioned weaknesses of knowledge transfer in unsupervised person re-identification, this thesis offers solutions as follow: **(1)** This thesis solves the insufficient of knowledge transfer by discovering transferable source knowledge at both the local *instance* and global *distribution* levels. This idea leads to a ***Hierarchical Unsupervised Domain Adaptation*** (HUDA) model. This is a non-trivial learning task due to the lack of direct correlations between source and target person identities. To solve this problem, we formulate a new cross-domain cross-class association learning algorithm.

(2) This thesis addresses the limitation of requiring extra target samples and narrow usage of the trained model by considering a "*train once, run everywhere*" pattern. In contrast to all the existing methods, the "*train once, run everywhere*" pattern considers a re-id model is trained by using the labelled data from a single *source* domain, and *frozen* it for universal deployment at any domains *without* further training and/or fine-tuning the model to any target domains (Figure

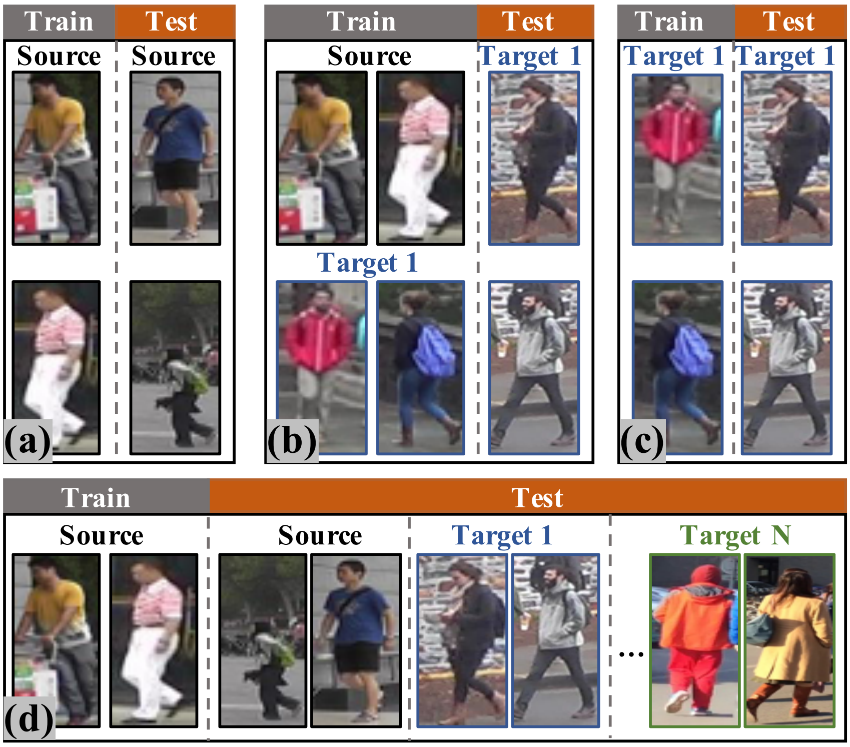


Figure 1.9: Learning strategies for person re-id: **(a)** *Supervised model learning* on a large set of cross-camera identity labelled training data per domain. Once trained, the model is deployed for the same domain alone. **(b)** *Unsupervised domain adaptation* on labelled training data from a source domain and unlabelled training data from the target domain. The adapted model is specific for the target domain. **(c)** *Unsupervised model learning* on unlabelled training data of the target domain. The trained model is specific for the target domain. **(d)** *Universal model learning* on labelled training samples from a source domain. Once trained the model can be *frozen forever* and applied for universal person re-id deployment at any target domains including the source domain. (Source Domain: Market-1501 dataset [43]; Target 1 Domain: DukeMTMC dataset [59]; Target N Domain: MSMT-17 dataset [41])

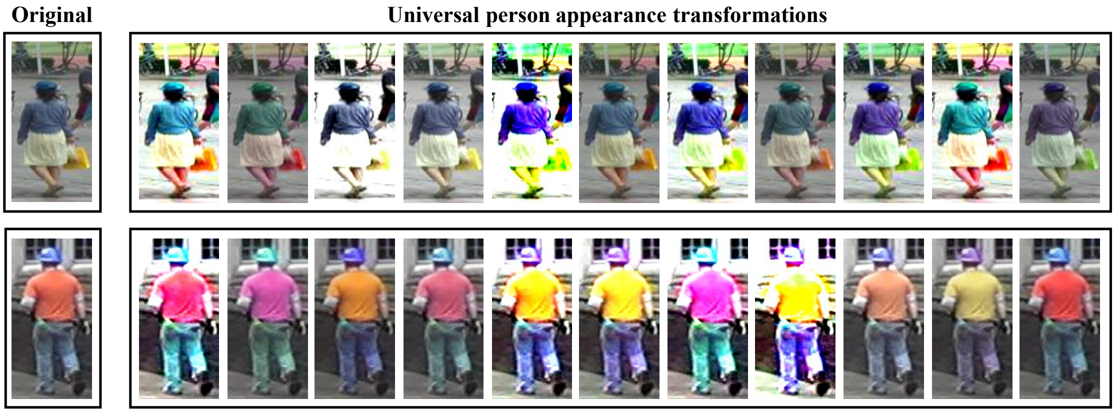


Figure 1.10: Universal person appearance transformations for training a single domain-generic re-id model enabling universal deployments. Compared to existing methods typically focusing on domain-specific model training (*train once, run once*), the proposed method allows for a *train once, run everywhere* pattern therefore favourably suits the industrial scale large system development without the need of training the system to every individual target domain as prior of each deployment.

1.9(d)). To this end, this thesis proposes a ***Universal Model Learning*** (UML) method capable of training a *single* model for domain generic person re-id deployment. UML trains a universally deployable re-id model *one-off* on transformed source training data, without the need of using any target domain data for model learning and refinement. The image transformations are designed to produce an extremely diverse training dataset (Figure 1.10) that simulates camera viewing condition variations as completely as possible for different domains, i.e. *domain complete therefore domain generic*. Viewing condition variations are simulated by randomly applying colour and contrast transformations to a labelled source person image. This image and its transformed versions share the same identity label. By design, the re-id model trained on the proposed augmented dataset is discriminative *universally* for any domains.

Assumptions. The proposed approach formulation is based on these assumptions :

1. Only considering global distribution in cross-domain knowledge transfer results in under-constrained optimisation problem, yielding suboptimal re-id models.

2. A large proportion of primitive attributes can be shared across domains in re-id, i.e. overlapped in the distribution.

3. Viewing condition variations are able to be simulated by randomly applying colour and contrast transformations to a labelled source person image.

1.3.3 Cross-model Knowledge Transfer on Image Classification

This thesis discusses the cross-model knowledge transfer on image classification in Chapter 5. In particular, knowledge distillation is studied as an example of cross-model knowledge transfer. Two issues of existing knowledge distillation methods are discussed and corresponding solutions are proposed.

- *Cross-model knowledge transfer requires significant extra computational cost and large memory (for a heavy teacher):* Whilst being able to create stronger target networks compared to the vanilla non-teacher based learning strategy, this scheme needs to train *additionally* a large teacher model with the expensive computational cost.

- *Cross-model knowledge transfer requires complex multi-phase training procedure and sing-phase training is sub-optimal due to lacking an appropriate teacher role:* While promising the student model quality improvement from aligning with a pre-trained teacher model, this strategy requires a longer training process, in a more complex multi-phase training procedure. These are commercially unattractive [60]. To simplify the distillation training process as above, simultaneous distillation algorithms [33, 60] have been developed to perform knowledge online teaching in a one-phase training procedure. Instead of pre-training a static teacher model, these methods train simultaneously a set of (typically two) student models which learn from each other in a peer-teaching manner. This approach merges the training processes of the teacher and student models, and uses the peer network to provide the teaching knowledge. Beyond the original understanding of distillation that requires the teacher model larger than the student, this online distilling strategy can improve the performance of any-capacity models, leading to a more generically applicable technique. Such a peer-teaching strategy sometimes even outperforms the teacher based offline distillation. The plausible reason is that the large teacher model tends to overfit the training data therefore leading to less extra knowledge on top of the manually labelled annotations [60]. However, the existing online distillation methods have a number of drawbacks: (1) Each peer-student model may only provide limited extra information, resulting in suboptimal distillation; (2) Training multiple students causes a significant increase of computational cost and resource burdens; (3) They require asynchronous model updating which has a notorious need of carefully ordering the operations of label prediction and gradient back-propagation across networks. We consider that all the weaknesses are due to the lack of an appropriate teacher role in the online distillation processing.

Solutions: To solve the aforementioned weaknesses, several approaches are developed as follows: **(1)** This thesis jointly solves both knowledge distillation for model compression and fast optimisation in model learning using a unified deep learning strategy. To that end, a ***Self-Referenced Deep Learning*** (SRDL) strategy is proposed that integrates the knowledge distillation concept into a vanilla network learning procedure. Compared to knowledge distillation, SRDL exploits different and *available* knowledge without the need for additionally training an expensive teacher by self-discovering knowledge with the target model itself during training. Specifically, SRDL begins with training the target network by a conventional supervised learning objective as a vanilla strategy, then extracts self-discovered knowledge (inter-class correlations) during model training, and continuously trains the model until convergence by satisfying two losses concurrently: a conventional supervised learning loss, and an *imitation loss* that regulates the classification probability predicted by the current (thus-far) model with the self-discovered knowledge. By doing so, the network learns significantly better than learning from a conventional supervised learning objective alone.

(2) This thesis provides a novel online knowledge distillation method that is not only more efficient (lower training cost) but also more effective (higher model generalisation improvement) as compared to previous alternative methods. In *training*, the proposed approach constructs a multi-branch variant of a given target network by adding auxiliary branches, creates a native ensemble teacher model from all branches on-the-fly, and learns simultaneously each branch plus the teacher model subject to the same target label constraints. Each branch is trained with two objective loss terms: a conventional softmax cross-entropy loss which matches with the ground-truth label distributions, and a distillation loss which aligns to the teacher's prediction distributions. In *test*, we simply convert the trained multi-branch model back to the original (single-branch) network architecture by removing the auxiliary branches, therefore *not* increasing test-time cost. In doing so, we derive an ***On-the-Fly Native Ensemble*** (ONE) teacher based simultaneous distillation training approach that not only eliminates the computationally expensive need for pre-training the teacher model in an isolated stage as the offline counterparts, but also further improves the quality of online distillation.

Assumptions. The formulation of proposed solutions is based on the following assumptions :

1. The vanilla training strategy relies only on the supervision of per-sample label, but ignores the discriminative knowledge incrementally discovered by the in-training model itself. It may lead to sub-optimal optimisation.

2. A multi-branch *single* model is more efficient to train whilst achieving superior generalisation performance and avoiding asynchronous model update.

1.3.4 Cross-task Knowledge Transfer on Image Classification

In Chapter 6, this thesis focuses on cross-task knowledge transfer. Recently, Few Shot Learning (FSL), especially in the context of image classification, attracts researchers attention, due to FSL aims to emulate human ability to quickly recognize the object in the new task. To learn a model which quickly adapts to the novel task, cross-task knowledge transfer is desire for FSL community. Here, two weaknesses in cross-task knowledge transfer are investigated in the FSL context.

- *The single prototype assumption of cross-task knowledge transfer in FSL is sensitive and unable to capture complex multi-modal class distribution (A class is represented as a set of clusters).* Existing FSL approaches fall into two main categories: (1) Metric learning based methods [61, 62, 63, 64] learn a distance metric between a query image and a set of annotated images such that the query image is closest to the annotated images of the same class; (2) Meta-gradient learning based methods [65, 66, 67, 68] focus on teaching a model to adapt quickly to new classes via a small number of regular gradient descent iterations. This thesis focuses on metric learning based methods due to their simplicity, flexibility and state of the art performance. The key idea of metric learning is to learn deep embeddings of input samples that minimises a pre-defined distance metric between samples of the same class. These methods typically rely on class prototypes, which are used to classify the unlabelled images via a nearest neighbour strategy. Prototypes can be defined as a global representation of a class that is calculated from the embedding of a set of annotated support images. Despite significant improvements achieved by metric learning approaches, existing metric based FSL approaches still suffer from an intrinsic drawback due to the general assumption that each category can be summarised using a

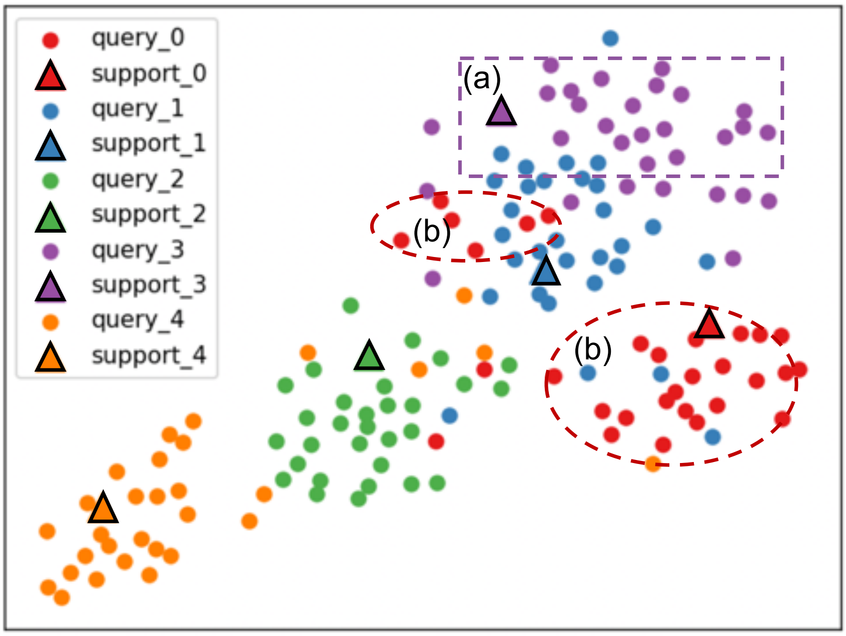


Figure 1.11: t-SNE visualization of feature embeddings from our baseline [69] for the support (triangles) and query images in the *miniImageNet* test stage under the 5-way 1-shot setting. This shows two drawbacks of single prototype metric learning methods: (a) the prototype can lack representative power and be out of distribution (purple rectangle); (b) a single prototype cannot accurately capture class multi-modal distributions (red ellipses).

single prototype. By only considering a uni-modal prototype per class, such methods are unable to capture complex multi-modal class distributions that often exist in real-world problems and fail to capture subtle differences between similar classes, as illustrated in Figure 1.11. Additionally, the performance of such models is very sensitive to the prototype quality. These two limitations motivate us to learn richer prototype representations that can capture latent data distributions accurately and enhance model robustness.

- *Cross-task knowledge transfer fails to exploit the unlabelled data and alternative Semi-Supervised Few-shot Learning (SS-FSL) solutions suffer from the error propagation.* SS-FSL investigates the benefit of incorporating unlabelled data in few-shot settings. Current state of the art SS-FSL [70, 71] methods rely on popular SSL techniques such as label propagation [72], that propagate label predictions to unlabelled data and self-training [73] that repeatedly labels unlabelled data, based on confidence scores, and retrains with the additional pseudo-annotated data. An important drawback of such strategies is their reliance on iteratively extending the training set using pseudo-label predictions. Building on pseudo-label decisions can propagate and amplify errors during training, yielding brittle methods sensitive to model initialisation and noisy data. This problem is exacerbated in few-shot scenarios, where available labelled data is highly limited and pseudo labels therefore have respectively larger influence.

Solutions: To solve these challenges, this thesis considers two different aspects:

(1) This thesis addresses the sensitive of cross-task knowledge transfer and the inability to capture complex multi-modal class by proposing a mixture of prototypes based metric learning approach. Relying on multiple prototypes was considered in [74, 75], proposing multiple prototype representations as clusters and local descriptors. Their promising approaches suffered from two important limitations: 1) prototypes were not optimised for diversity [74], limiting the benefits of the multiple representations and 2) local descriptors were not regularized [75], yielding prototypes of potential poor representative power due to the use of local inputs. Our key idea is to learn a set of prototypes *per class* that are optimised to maximise individual (high representative power) and ensembling performance (high inter-prototype *variance*). This is achieved by computing a set of local and global class prototypes, which allows to focus on different regions and image attributes. We regularize local prototypes with a) a soft attention gate to merge prototype classification decisions, effectively allowing unreliable and non-discriminative prototypes (image regions) to be ignored and b) a self-supervised task that regularizes the learning process on local inputs, yielding robust and class-representative local prototypes. This approach allows us to separate and generalise to new classes accurately due to the resulting richer representations. Moreover, the increased robustness granted by our mixture of prototypes allows to use an imprinted weights formulation [69] and maintain high performance, while alleviating the requirement to retrain a model when new categories are available.

(2) In light of these limitations of explored unlabelled data, this thesis deviates from an iterative pseudo-labelling scheme and considers an alternative SSL strategy, relying instead on the concept of Consistency Regularisation (CR) [76, 77, 78, 79]. The approach successfully exploits unlabelled data without introducing pseudo-label requirements and outperforms preexisting state of the art methods. Applying CR to the SS-FSL setting is however non-trivial in low-data regimes. The proposed two-stage approach draws on both the concepts of imprinted class prototypes [69] and CR [76, 79]. It comprises semi-supervised pre-training on base classes, followed by Dynamic Prototype Refinement (DPR) on novel classes. The imprinted weights approach allows the learning of class prototype representations as model weights using standard end-to-end training (*vs*. commonly used episode training) via the use of a cosine classifier. This allows to seamlessly introduce a CR task in the base class training process, effectively leveraging unlabelled data. The CR task borrows ideas from [79], and is formulated within a mean teacher framework [76] using a weak-strong augmentation strategy, where the prediction of highly perturbed inputs must match those of the same, weakly perturbed, input image. This base class training stage allows high quality initial prototypes to be directly inferred from labelled image features, providing a robust initialisation for novel classes. The novel DPR stage exploits unlabelled samples towards learning prototypes of higher quality. DPR approach alternates between explicitly updating of prototypes using unlabelled samples that yield the most confident predictions (*i.e*. nearest to their assigned class prototype), and implicitly fine-tuning the model with CR on a second selection of unlabelled samples. Alternating between typically smaller, more conservative updates (implicit refinement) and larger, often times more disruptive pseudo-label based updates (explicit refinement), results in faster convergence for CR and often large performance gains, whilst at the same time affording robustness to pseudo-labelling errors.

Assumptions. The proposed approach formulation is based on these assumptions :

1. Prototypes can be defined as a global representation of a class that is calculated from the embedding of a set of annotated support images.

2. Only considering a uni-modal prototype per class in FSL, such methods are unable to capture complex multi-modal class distributions that often exist in real-world problems and fail to capture subtle differences between similar classes and the performance of such models is very sensitive to the quality of the prototypes.

3. CR currently fails in the SS-FSL scenario due to 1) the slow convergence of CR techniques [79], which is in conflict with FSL fast convergence requirements, in order to alleviate overfitting risks and 2) the poor reliability of teacher predictions in early stages when trained on limited data.

1.4 Contributions

The contributions made in this thesis are summarised below:

Chapter 3 considers the problem of optimising attention selection and multi-scale matching within any auto-detected person bounding boxes in cross-class knowledge transfer for maximising performance on re-id tasks. To optimise attention selection scheme, this chapter formulates a novel **I**dentity **D**iscriminativ**E** **A**ttention reinforcement **L**earning (IDEAL) model for attention selection *post-detection* given re-id discriminative constraints. Extensive experiments on two large auto-detected datasets CUHK03 [80] and Market-1501 [43] demonstrate the advantages of the proposed IDEAL model over a wide range of contemporary and state-of-the-art person re-id methods.

Besides, this chapter addresses the multi-scale person search challenge by proposing a Cross-Level Semantic Alignment (CLSA) deep learning approach capable of learning more discriminative identity feature representations in a unified end-to-end model. Extensive experiments show the modelling advantages and performance superiority of CLSA over the state-of-the-art person search and multi-scale matching methods on two large person search benchmarking datasets: CUHK-SYSU and PRW.

Chapter 4 proposes a novel idea of exploring instance-wise localised source knowledge to enhance the cross-domain knowledge transfer for unsupervised person re-id. It addresses the limitations of existing global feature distribution adaptation based methods for cross-domain knowledge transfer. This chapter further formulates a Hierarchical Unsupervised Domain Adaptation (HUDA) method. Extensive evaluations demonstrate the superiority of HUDA over a variety of state-of-the-art models for unsupervised cross-domain person re-id on four benchmarks: Market-1501 [43], DukeMTMC [59, 48], MSMT17 [41], and CUHK03 [40].

Furthermore, this chapter presents a “train once, run everywhere” pattern for general cross-domain knowledge transfer for universal person re-identification. This is the first deep learning attempt of universal person re-id. This chapter proposes a simple yet effective Universal Model Learning (UML) approach for realising universal person re-id. Extensive evaluations demonstrate the model training and performance superiority of UML over the state-of-the-art alternative methods on five person re-id benchmarks: Market-1501, DukeMTMC-reID, CUHK03, MSMT17, and VIPeR.

Chapter 5 studies the expensive computation cost of acquiring knowledge to do cross-model knowledge transfer and addresses the drawback of multi-stages and multi-models in knowledge distillation. This chapter investigates for the first time the problems of knowledge distillation based model compression and fast optimisation in model training using a unified deep learning approach, an under-studied problem although both problems have been studied independently in the literature. Specifically, this chapter presents a stage-complete learning rate decay schedule in order to maximise the quality of intermediate self-discovered knowledge and therefore avoids the negative guidance to the subsequent second-stage model optimisation. Besides, this chapter further introduces a random model restart scheme for the second-stage training with the purpose of breaking the optimisation search space constraints tied to the self-referenced deep learning process. Extensive comparative experiments are conducted on object categorisation tasks (CIFAR10/100 [81], Tiny ImageNet [82], and ImageNet [3]) and person instance identification tasks (Market-1501 [43]).

To alleviate the weakness of the multi-stage and multi-model training required by the knowledge distillation. This chapter proposes a novel online knowledge distillation method (ONE) that is not only more efficient (lower training cost) but also more effective (higher model generalisation improvement) as compared to previous alternative methods. Extensive experiments on four benchmarks (CIFAR10/100, SVHN, and ImageNet) show that the proposed ONE distillation method enables to train more generalisable target models in a one-phase process than the alternative strategies of offline learning a larger teacher network or simultaneously distilling peer students, the previous state-of-the-art techniques for training small target models.

Chapter 6 discusses the cross-task knowledge transfer in few shot learning under the context of image classification. This chapter presents a Mixture of Prototypes (MP) learning strategy for metric-based FSL. A simple and generic approach that can easily be embedded in popular metric learning based methods. Extensive experiments demonstrate the superiority of our method compared to the state-of-art on two standard benchmarks: *mini*ImageNet and *tiered*ImageNet.

To further utilize and transfer the knowledge of the unlabelled data, this chapter presents “Fewmatch”; a novel semi-supervised few-shot learning approach that alleviates the need for iterative pseudo-labelling. This is the first approach exploiting the power of consistency regularisation in an SS-FSL context. This chapter further introduces the concept of dynamic prototype refinement. By iteratively updating prototypes for novel categories, using both explicit feature averaging and implicit fine-tuning through a two-level top-K selection scheme, Fewmatch is able to successfully leverage CR in a few-shot setting and are robust to error propagation inherent to pseudo-labelling schemes. Extensive experiments demonstrate that we achieve state of the art performance on two standard benchmarks.

1.5 Thesis outline

The remaining chapters of this thesis are organised as follows, with all chapters are structured in Figure 1.12:

Chapter 2 provides a literature review of fine/coarse-grained object recognition by deep learning. Furthermore, a set of existing cross-class/domain/model/task knowledge transfer approaches are discussed, including knowledge distillation, domain adaption, few shot learning.

Chapter 3 investigates the inaccurate person bounding box and multi-scale challenge of cross-class knowledge transfer for fine-grained person re-id. Specifically, to refine of inaccurate person bounding box, this chapter proposed a novel **I**dentity **D**iscriminativ**E** **A**ttention reinforcement **L**earning **(IDEAL)** model for attention selection *post-detection* given re-id discriminative constraints. Furthermore, this chapter discusses the multi-scale problem in both person re-id and person search, proposing a Cross-Level Semantic Alignment (**CLSA**) model to address the multi-scale matching challenge.

Chapter 4 tackles the insufficient cross-domain knowledge transfer in Unsupervised Domain Adaptation and unscalable for practical large scale models. This thesis addresses this insufficient knowledge transfer by discovering transferable source knowledge at both the local *instance* and global *distribution* levels. This idea leads to a Hierarchical Unsupervised Domain Adaptation (**HUDA**) model. To increase the usage and scalability of existing UDA methods on real-world deployment. This chapter further proposes a "*train once, run once*" pattern by presenting a Universal Model Learning (**UML**) method, enabling training a single model for domain generic deployment.

Chapter 5 discusses extra computation cost and complex multi-phase training of cross-model knowledge transfer for coarse-grained object recognition. Specifically, this chapter proposes two alternative methods to solve the above limitations. (1) A Self-Referenced Deep Learning (**SRDL**) strategy is proposed to combine the fast optimization in knowledge distillation to solve the large training cost required by pre-trained teacher network. (2) This chapter designs an On-the-Fly Native Ensemble (**ONE**) teacher based simultaneous distillation training approach that not only eliminates the computationally expensive need for pre-training the teacher model in an isolated stage as the offline counterparts, but also further improves the quality of online distillation.

Chapter 6 provides the cross-task knowledge transfer in the few shot classification under supervise and semi-supervised context, respectively. For conventional supervised few shot learning, this chapter proposes a generic Mixture of Prototypes (**MP**) learning strategy that can be embedded in popular metric learning based methods and improve their performance. MP is tailored to compute richer and more robust representations in contrast to conventional single prototype approaches. To facilitate the knowledge transfer through unlabelled data, this chapter further presents "**Fewmatch**"; a novel semi-supervised few-shot learning approach that alleviates the need for iterative pseudo-labelling. The proposed prototype-based consistency regularization model effectively leverages unlabelled samples in order to obtain reliable prototype-based predictions. Furthermore, this chapter introduces the concept of dynamic prototype refinement. By iteratively updating prototypes for novel categories, using both explicit feature averaging and implicit fine-tuning through a two-level top-K selection scheme, we are able to successfully leverage CR in a few-shot setting and are robust to error propagation inherent to pseudo-labelling schemes

Chapter 7 offers the conclusion and suggests the future work related to this thesis content.

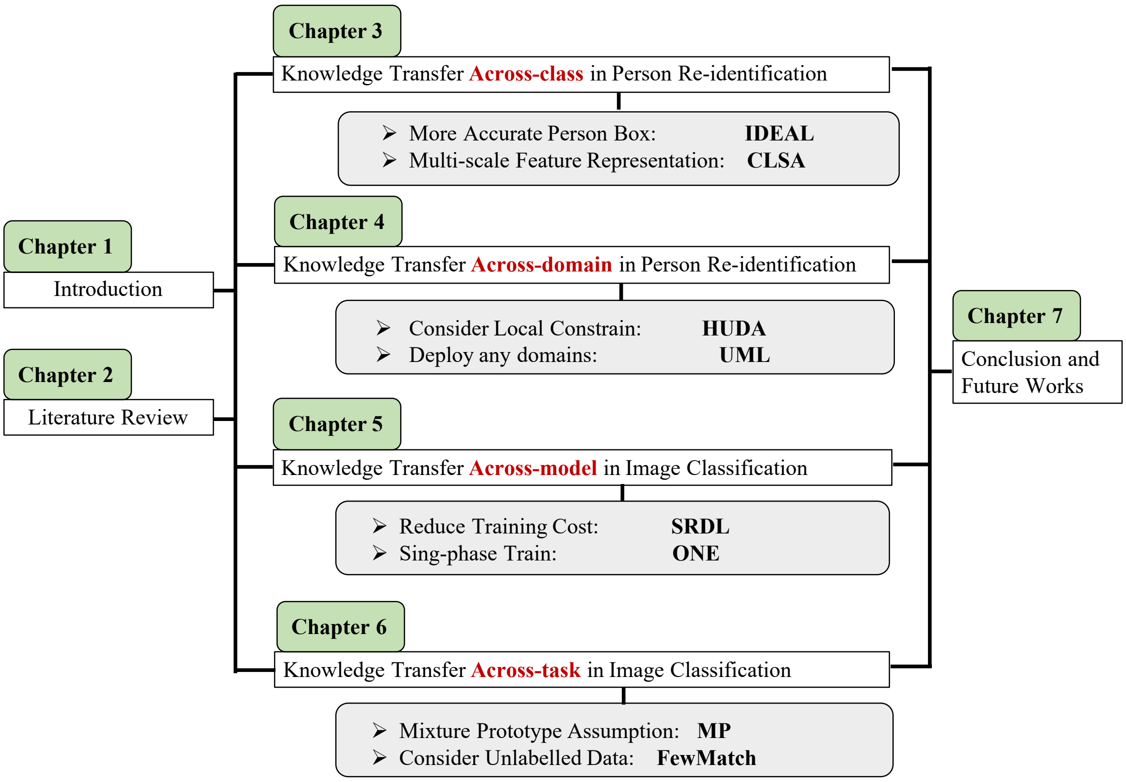

Chapter 3
Knowledge Transfer Across-class in Person Re-identification
More Accurate Person Box: IDEAL
Multi-scale Feature Representation: CLSA
Chapter 4
Knowledge Transfer Across-domain in Person Re-identification
Consider Local Constrain: HUDA
Deploy any domains: UML
Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 7
Conclusion and Future Works
Chapter 5
Knowledge Transfer Across-model in Image Classification
Reduce Training Cost: SRDL
Sing-phase Train: ONE
Chapter 6
Knowledge Transfer Across-task in Image Classification
Mixture Prototype Assumption: MP
Consider Unlabelled Data: FewMatch

Figure 1.12: Summarisation and structure of all chapters.

Chapter 2

Literature Review

In the long history of humankind (and animal kind too) those who learned to collaborate and improvise most effectively have prevailed.

—— **Charles Darwin**

This chapter gives a rough summary on existing approaches about knowledge transfer in object recognition which are closely related to this thesis. Section 2.1 discusses the development of object recognition systems under the wave of deep learning. Specifically, the techniques on the task of fine-grained person re-identification and coarse-grained image classification are reviewed. Section 2.2 focuses on knowledge transfer mechanism of visual recognition in the literature. Section 2.3 offers the preliminaries of deep learning for object recognition, including essential mathematical formulation and widely used object function. Finally, this chapter provides a summary of the literature review on Section 2.4.

2.1 Deep Learning for Object Recognition

2.1.1 Image Classification

Visual recognition, especially, image classification is one of the crucial tasks in computer vision and it refers to distinguishing images in different categories based on their semantic meaning. Despite the simplicity of definition, image classification has wide applications in reality such as face recognition, video understanding in surveillance system, traffic scene recognition in transportation systems, content-based image retrieval and image classification in medicine industry.

In image classification, feature selection is an important aspect. Earlier work, including HOG [83] and SIFT [84] have been widely used to extract features. However, such hand-engineered features suffer from limited representation power and more importantly lack semantic and abstract concept understanding, therefore having limited generality. In contrast, recently deep convolutional neural network (CNN) architectures achieved superior performances due to their strong feature representation ability and potential to capture high-level abstract features. CNN networks often contain a set of sequentially connected linear convolution layers with non-linear activated function followed and an additional fully connected layer for prediction in the last. The stacking of multiple linear and non-linear processing units in a layer-wise fashion enables the model to learn complex representations at different levels of abstraction. Consequently, in some complex recognition tasks, *i.e.* ImageNet [3] consisting of one thousand image categories, deep CNNs have shown substantial performance improvements over conventional hand-engineered feature methods [83, 84].

A few years ago, researchers found the representational capacity of CNN can be increased by designing more complex architecture when sufficient training data is available [16]. After that, several attempts such as modification of processing units, parameter and hyper-parameter optimization strategies, design patterns have applied to the deep CNN, aiming to increase CNN capacity and representation power, which enables CNN scalable to large complex multi-classes problems. The feature extracted from the CNN network rather than the hand-engineered became prevalent after the exemplary performance of AlexNet on the ImageNet dataset in 2012 [16]. Following their work, Simonyan and Zisserman introduce the VGG net [4], with the principle of a simple and homogenous topology. After that, inception block [9] is proposed with the core idea which is a split, transform and merge. This is the first attempt to construct branches within a layer, which allows the same level of features at different spatial scales. To enable convergence and well-training of the extremely deeper network, skip connection is presented in ResNet [5], resulting in more than 1000 layers of architecture. Afterwards, the concept of skip connection is widely applied on most of the succeeding networks, such as Inception-ResNet [85], Wide ResNet [13], ResNeXt [86], DenseNet [8], etc.

To reduce these onerous development costs in architecture engineering, Neural Architecture

Search (NAS) approaches are proposed to search for the best architecture in a variety of tasks, *i.e.* object recognition. The main goal of NAS is to produce a robust and well-performing neural architecture by selecting and combining different basic components. Existing NAS methods can be roughly divided into two categories 1) model structure type [87, 88, 89, 90, 91, 92]. 2) model structure design by hyperparameter optimization (HPO) [93, 94, 95, 96, 97, 98].

While existing deep neural architectures found by either human design or NAS are able to achieve super-human performance, most of them are trained independently without considering the cross-model knowledge transfer. This thesis provides solutions to alleviate this issue in Section 5

2.1.2 Person Re-id and Search

Person re-identification (re-id) aims at searching people across non-overlapping camera views distributed at different locations by matching person bounding box images [6]. The nature of non-overlap person identities between training and testing requires highly demand of cross-class knowledge transfer. This section gives the literature review on the direction of cross-class knowledge transfer in re-id, and other related aspects such as cross-domain knowledge transfer are provided in the next section.

Most existing re-id methods [99, 100, 101, 102, 103, 104, 35, 105, 80, 106, 107, 42, 108] focus on supervised learning of person identity-discriminative information. Representative learning algorithms include ranking by pairwise or list-wise constraints [107, 109, 105, 110], discriminative subspace/distance metric learning [100, 101, 102, 104, 42, 108, 111], and deep learning [112, 106, 80, 113, 29, 106, 114]. They typically require a large quantity of person bounding boxes and inter-camera pairwise identity labels, which is prohibitively expensive to collect manually. In parallel to supervised learning re-id methods, unsupervised learning methods have started to gain increasing potentials for eliminating the need for labelling large training data [115, 116, 56, 117, 118]. Such methods rely on the reconstruction loss designs [115, 116] or self-discovered cross camera label information by the in-training model for self-supervised learning [56, 117, 118].

Existing re-id methods often assume the provided person bounding box is accurate and in the same resolution. However, this is unrealistic in the real-world deployment and yields the negative influence of cross-class knowledge transfer. To address these issues, there are several attempts, including automatic detection, saliency and attention selection, person search and multi-scale match.

Automatic Detection in Re-id

In real-world re-id scenarios, automatic person detection [119] is essential for re-id to scale up to large size data, e.g. re-id benchmarks CUHK03 [80] and Market-1501 [43]. Most existing re-id test datasets (Table 2.1) are *manually cropped*, as in VIPeR [120] and iLIDS [121], thus they do not fully address the re-id challenge in practice. However, auto-detected bounding boxes are not optimised for re-id tasks, due to potentially more background clutter, occlusion, missing body part, and inaccurate bounding box alignment (Figure 1.7). This is evident from that the rank-1 re-id rate on CUHK03 drops significantly from 61.6% on manually-cropped to 53.4% on auto-detected bounding boxes by state-of-the-art hand-crafted models [122], that is, an 8.2% rank-1 drop; and from 75.3% on manually-cropped [29] to 68.1% on auto-detected [123] by state-of-the-art deep learning models, that is, a 7.2% rank-1 drop. Moreover, currently reported "auto-detected" re-id performances on both CUHK03 and Market-1501 have further benefited from artifical *human-in-the-loop cleaning process*, which discarded "bad" detections with $< 50\%$ IOU (intersection over union) overlap with corresponding manually cropped bounding boxes. Poorer detection bounding boxes are considered as "distractors" in Market-1501 and not given re-id labelled data for model learning. Recent works [80, 43, 124, 124] have started to use automatic person detection for re-id benchmark training and test. Auto-detected person bounding boxes contain more noisy background and occlusions with misaligned person cropping (Figure 1.7), impeding discriminative re-id model learning.

There is very little attempt in the literature for solving this problem of attention selection within auto-detected bounding boxes for optimising person re-id, except a related recent study on joint learning of person detection and re-id [125]. The proposed approach IDEAL (Section 3.1) however differs from that by operating on any third party detectors *independently* so to benefit continuously from a wide range of detectors being rapidly developed by the wider community. Other related possible strategies include local patch calibration for mitigating misalignment in pairwise image matching [126, 127, 128, 123] and local saliency learning for region soft-selective matching [129, 130, 126, 131]. These methods have shown to reduce the effects from viewpoint and human pose change on re-id accuracy. However, *all* of them assume that person bounding boxes are reasonably accurate.

Saliency and Attention Selection in Re-ID

Most related re-id techniques are localised patch matching [127, 128, 123] and saliency detection [129, 130, 126, 131]. They are inherently unsuitable by design to cope with poorly detected person images, due to their stringent requirement of tight bounding boxes around the whole person. In contrast, the proposed IDEAL model (Section 3.1) is designed precisely to overcome inaccurate bounding boxes therefore can potentially benefit all these existing methods.

Table 2.1: Person re-id datasets with/without auto-detection. MC: Manual Cropping; AD: Automatic Detection.

Dataset	VIPeR [120]	GRID [132]	iLIDS [121]	CAVIAR4ReID [47]	CUHK03 [80]	Market-1501 [43]
Year	2007	2009	2010	2011	2014	2015
Annotation	MC	MC	MC	MC	MC+AD	AD
Identities	632	250	119	72	1,360	1,501
Images	1,264	1,275	476	1,221	28,192	32,668

Person Search

Person search aims to find a probe person in a gallery of whole unconstrained scene images [28]. It is an extended form of person re-id [6] by additionally considering the requirement of automatically detecting people in the scene images besides matching the identity classes. Unlike the conventional person re-id problem assuming the gallery images as either manually cropped or carefully filtered auto-detected bounding boxes [29, 20, 30, 31, 32, 33, 34, 35, 36, 33], person search deals with raw unrefined detections with many false cropping and unknown degrees of misalignment. This yields a more challenging matching problem especially in the process of person re-id.

In the literature, there are only a handful of person search works [28, 37, 38]. Xiao et al. [28] propose a joint detection and re-id deep learning model for seeking their complementary benefits. Zheng et al. [37] study the effect of person detection on the identity matching performance. Liu et al. [38] consider recursively search refinement to more accurately locate the target person in the scene. While existing methods focus on detection enhancement, we show that by a state-of-the-art deep learning object detector with small improvements, person localisation is not a big limitation. Instead, the multi-scale matching problem turns out a more severe challenge in person search (Section 3.2). In other words, solving the multi-scale problem is likely to bring more performance gain than improving person detection (Figure 3.8(c)).

Multi-scale Match in Re-identification

Given the manual construction nature of re-id datasets, the scale diversity of gallery images tends to be restricted, compared to the person search benchmarks. It is simply harder for humans to verify and label the person identity of small bounding boxes, therefore leading to the selection and labelling bias towards large boxes (Figure 1.8(b)). Consequently, the intrinsic multi-scale matching challenge is *artificially* suppressed in re-id benchmarks, hence losing the opportunity to test the real-world model robustness. Existing re-id methods can mostly afford to ignore the problem of multi-scale person bounding boxes in algorithm design. Whilst extensive efforts have be made to solving the re-id problem [35, 40, 133, 134, 29, 20, 30, 43, 32, 135, 36, 136, 34, 56, 53, 137, 138], there are only limited works considering multi-scale matching [137, 139]. Liu et al. [139] firstly developed a multi-scale triplet deep architecture, aiming to learn a combination of different scale features for given a person bounding box. which learns deep features of a pedestrian at different scales. In particular the architecture integrates both shadow and deep networks, yielding low-level and high-level appearance features from images, respectively. Chen et al. [137] proposed multi-scale person features learning model by Convolutional Neural Networks (CNN), aiming to jointly learn discriminative scale-specific features and maximise multi-scale feature fusion selections in image pyramid inputs. Specifically, they formulate a novel Deep Pyramid Feature Learning (DPFL) CNN architecture for multi-scale appearance feature fusion optimised by concurrent per-scale re-id identity losses and interactive cross-scale consensus regularisation.

Beyond all these existing methods, CLSA introduced in this thesis (Section 3.2) is designed specifically to explore the in-network feature pyramid in deep learning for more effectively solving the under-studied multi-scale challenge in both person re-id and search.

2.2 Knowledge Transfer

2.2.1 Domain Adaptation

Domain adaptation is a field associated with cross-domain knowledge transfer. This scenario arises when we aim at applying an algorithm trained in one or more source domain to a different (but related) target domain. The most commonly used domain adaptation approaches can be roughly classified into two categories [45]: 1) feature adaptation 2) image adaptation. The former transfers the discriminative feature information learned from the labelled source training data to the target feature space by distribution alignment. These methods often use discrete attribute labels for facilitating the information transition across domains due to their better domain invariance property than low-level feature representations. On the contrary, the latter aims to transfer the labelled source identity classes from the source domain to the target domain through cross-domain conditional image generation in the appearance style and background context at the pixel level. The synthetic images are then used to fine-tune the model towards the target domain.

The most typical feature adaptation methods can be found in [140, 141, 142, 143]. They are usually motivated by minimizing the distribution discrepancy between the source and target domain in the share feature space. Specifically, Tzeng et al. [140] and Long et al. [141, 14] minimise Maximum Mean Discrepancy (MMD) metric to align the global distribution between source and target. Another popular metric is in deep CORAL [142], they minimise the difference of the feature covariance matrice between the source and target domain instead of MMD metric. Image adaptation approaches based on generative adversarial methods [144, 145, 146] perform similar distribution alignment in raw pixel space rather than feature space. They usually focus on learning a generator network to translate source data to the "style" of a target domain.

Beyond the above mentioned Domain Adaptation (DA) approaches on the general image classification task, this thesis focuses on DA in a more challenging person re-id tasks.

Unsupervised Domain Adaptation for Person Re-id

Most existing person re-id methods require *supervised* learning on a large labelled training dataset [43, 137, 42, 34, 39, 44]. They assume that the training and test data are sampled from the same domain and have limited cross-domain generalisation, therefore suffering the intrinsic domain gap problem. As a result, they have poor scalability to large scale re-id deployments in real-world when a large labelled training set is unavailable. While reducing the labelling effort, semi-supervised learning [147, 148] still needs some cross-camera pairwise labels which may not be available inherently. A straightforward solution is to use unsupervised learning that exploits less discriminative hand-crafted features. Besides, hand-crafted features [46, 47, 149, 116, 150, 52] are largely domain-generic and suffer from significantly weaker re-id matching performance. Recently, unsupervised domain adaptation (UDA) methods have demonstrated increasing significance in solving cross-domain re-id deployments [49, 50, 41, 53, 54]. The existing UDA models fall into two categories: (1) image adaptation (synthesis) [49, 50, 151, 152, 153], and (2) feature adaptation [52, 154, 155, 53, 54]. The *first* approach is often built on Generative Adversarial Networks (GANs) [156]. The main idea is to transform the labelled source domain images into the style of the unlabelled target domain while attempting to preserve the person identity information. In doing so, the source class labels can be used for supervised learning on the synthetic imagery. The *second* approach adopts a global feature distribution alignment strategy. This assumes that the model discrimination is related to global feature distribution statistics. Conceptually, both approaches are based on global data distribution alignment, with the key difference in its data form (image pixels or feature dimensions). One of their common weaknesses is that they all suffer from a *highly under-constrained learning* problem. That is, both do not consider instance level alignment to enable explicit fine-grained source knowledge adaptation. The proposed HUDA (Section 4.1) addresses this limitation by formulating a unified model for simultaneous global (distribution alignment) and local (instance alignment) knowledge transfer and adaptation across domains. Experiments show clearly the added benefits from modelling both levels of knowledge adaptation in order to maximise cross-domain knowledge transfer between the labelled source and the unlabelled target domains. In comparison to UDA, unsupervised deep learning [56] provides an *orthogonal* strategy. It aims to self-mine re-id discriminative information from the unlabelled training data in the target domain. This is typically done by feature learning. analogous to the component of hybrid approach [50] involving unsupervised feature learning of target training data. It is generally beneficial to model performance by combining different strategies, for instance, integrating feature adaptation with image generation [50, 57] or unsupervised learning [58]. This thesis further evaluates HUDA integrated with unsupervised learning.

Hybrid Learning Person Re-id.

Recently, some methods [57, 157] consider a hybrid of domain adaption and unsupervised learning approach. Their basic idea is exploiting the image synthesis/feature alignment to transfer knowledge from source to target, while using unsupervised learning method on unlabelled target data to constrain the re-id feature learning at the same time.

Universal Learning Person Re-id.

In this thesis, Section 4.2 presents a universal learning person re-id approach (UML), which differs dramatically from all the existing methods as discussed above. UML trains a single *domain-generic* re-id model for universal deployments. This is in contrast to previous learning algorithms usually producing *domain-specific* models using either labelled source and/or unlabelled target domain training data. That being said, a separate model training is required for each target domain deployment which neither is cost-effective and convenient nor always allowed for industrial settings. The re-id model trained by UML can be immediately deployed to any domains where no video and image data are observed to model optimisation. Such universal deployment property is favourable and desired to practical system development. Moreover, the proposed image transformation method is computationally efficient with flexible design due to no need for complex model formulation and costly pixel synthesis model training. In comparison to hand-crafted features [158, 46, 42, 47, 43], UML model has the extra capability for feature representation learning and model optimisation as supervised and unsupervised learning counterparts, whilst simultaneously retaining the merits of domain universality as hand-crafted features. Besides, UML learning method differs from and is more scalable than multi-target domain adaptation [159, 160] where all target domains need to be seen to training. Conceptually, UML generalises the notion of multi-target domain simultaneous adaptation since we make a model effective for all different domains even without accessing any target domain data.

2.2.2 Knowledge Distillation

There are a number of attempts at transferring knowledge between varying-capacity networks [161, 25, 26, 27]. One of the potential directions is knowledge distillation. A representative work is introduced by [25], and they successfully used the well trained large network to help to train the small network by knowledge distillation. The rationale behind is taking advantage of extra supervision provided by the teacher model during training the target model, beyond a conventional supervised learning objective such as the cross-entropy loss subject to the training data labels. The extra supervision is typically extracted from a pre-trained powerful teacher model in form of class posterior probabilities [25], feature representations [26, 27], or inter-layer flow (the inner product of feature maps) [162]. Recently, some theoretical analysis has been provided to relate distillation to information learning theory for which a teacher provides privileged information (e.g. sample explanation) to a student in order to facilitate fast learning [163, 164]. Zhang et al. [165] exploited this idea for video based action recognition by considering the computationally expensive optic flow as privileged information to enhance the learning of a less discriminative motion vector model. This avoids the high cost of computing optic flow in model deployment whilst computing cheaper motion vectors enables real-time performance.

In contrast to all the above existing works, this thesis provides Self-Reference Deep Learning (SRDL) in Section 5.1, aiming to eliminate the extra expensive teacher model training cost.

To this end, SRDL uniquely explores self-discovered knowledge in target model training by *self-distillation*, therefore more cost-effective. Concurrent with SRDL, Furlanello et al. [166] independently propose training the networks in generations, in which the next generation is jointly guided by the standard one-hot classification labels and the knowledge learned in the previous generation. However, the training budget of each generation is almost the same as the vanilla strategy, leading to the total cost of this method several times more expensive than vanilla training.

Online Distillation

Earlier distillation methods often take an offline learning strategy, requiring at least two phases of training. The more recently proposed deep mutual learning [33] overcomes this limitation by conducting an online distillation in one-phase training between two peer student models. Anil et al. [60] further extended this idea to accelerate the training of large scale distributed neural networks. and show its promising on the industrial community.

However, the existing online distillation methods lack a strong "teacher" model which limits the efficacy of knowledge discovery. As the offline counterpart, multiple nets are needed to be trained therefore computationally expensive. This thesis overcomes both limitations in Section 5.2 by designing a new online distillation training algorithm characterised by simultaneously learning a teacher on-the-fly (ONE) and the target network as well as performing batch-wise knowledge transfer in a one-phase training procedure.

2.2.3 Few Shot Learning

Few shot learning (FSL) involves training a model towards emulating the human ability to perform a novel task well (*e.g.* recognise new categories) using only very few examples. Therefore, the FSL task requires a high demand of cross-task knowledge transfer. Recent interest and activity in FSL have resulted in a rapidly growing, rich body of work. Here, this section briefly highlights the most relevant approaches in the context of classification to the thesis.

Supervised Few Shot Learning

Existing Supervised FSL approaches can be broadly divided into two categories: (1) Meta-gradient learning; (2) Metric-learning;

Meta-gradient learning can be regarded as a first pillar of FSL approaches. Many recent works train a meta-learner using the *learning-to-learn* paradigm [167, 168, 169, 68]. A popular strategy within this paradigm involves finding optimal network parameter initializations [65, 170, 171, 172, 66, 67] such that fine-tuning becomes fast and requires only a few weight updates.

Metric-learning based techniques constitute the second branch of FSL methods. A distance metric between a query image and a set of labelled images is learned such that the query image is closest to labelled images of the same class [173, 62, 61, 174, 63]. The crux of metric learning involves learning a good *proxy* per class that is used to classify unlabelled images at test time, typically with a nearest neighbour strategy. The common approach for defining the representative class proxy involves using the average feature representation of a set of labelled images. Recent parameter generating methods [64, 175] alternatively propose to generate proxies using classifier weights. At training time, a cosine classifier is learned on top of feature extraction layers and each column of classifier parameter weights can be regarded as a proxy for the respective class. At test time, a new class proxy (new column of classifier weight parameters) is defined by averaging the feature representation of support images, similar to previous metric learning approaches. Inspired by the recent progress of self-supervised learning [176, 177, 178], the work of [179] alternatively strengthens image level representations using self-supervision for auxiliary task learning. The approach proves beneficial for learning generalizable feature representations in single proxy models. Recent metric-learning FSL methods have proved highly successful, however common drawbacks that remain are caused in part by the limitation of relying on a single proxy representation per class (Figure 1.11).

In contrast to learning image level representations, recent work [75] presents evidence that global image based measures may be too coarse to be effective in few-shot scenarios, where samples are scarce. The authors instead propose to learn local descriptors for their image-to-class measure. Allen et al. [74] alternatively propose Infinite Mixture Prototypes (IMP). The IMP approach represents each class as a set of clusters (prototypes), each consisting of class image representations. The strategy of increasing model capability beyond single proxy approaches is similar in spirit to our approach. However, tackling class representation with the IMP clustering strategy does not afford any mechanism to account for prototype diversity. Furthermore, in contrast to IMP and [75], the proposed methods MP in Section 6.1 combines image level representations with local descriptors and carefully regularise local proxy influence using self supervision and attention in order to maximise proxy diversity and representative power. Finally, MP is the sole approach that is designed in a generic way and can be integrated into pre-existing metric-based FSL methods.

Semi-supervised Few Shot Learning

Semi-supervised Learning (SSL) The objective of SSL is to use latent information, provided by additional unlabeled data, towards improving the quality of decision boundaries with respect to an underlying data distribution. Existing SSL methods generally fall into two categories: (1) Pseudo-labelling and (2) Consistency Regularisation. Techniques in the former category iteratively assign pseudo labels to the unlabelled samples such that they can then be used with a supervised loss. These include directly using the network class prediction [73] and graph-based label propagation [72]. A number of SSL works build on the second category of Consistency Regularisation [180, 77, 76], and have achieved impressive results. The crux of the idea of CR is to encourage invariant (stable) predictions for a given sample under different perturbations towards improving class decision boundaries. CR ideas were first explored in [180, 77] and extended in [76] where the authors propose a mean teacher framework, to perform CR between a student and teacher model in a learning paradigm where models share the same architecture and teacher parameters are updated as an exponential moving average of the student weights. Several works such as ICT [181], Mixmatch [78] and Remixmatch [79] have then enabled sample perturbations by creating variants of mixup samples [182] that can then be further perturbed. Encouraged by the benefits that result from representing class information using prototypes [62, 69], we take an alternative approach to CR in the context of SS-FSL and influence model prediction by considering a measure of distance between unlabelled data and class prototypes.

Semi-Supervised Few-Shot Learning (SS-FSL) Existing SS-FSL approaches are based on the pseudo-labelling strategy that was discussed in the context of SSL. Ren *et al.* [183] propose mask soft K-means, based on the metric learning approach, ProtoNets [62]. The authors use a soft K-means and iteratively assign pseudo labels to tune prototypes. More recently [70] propose a Transductive Propagation Network (TPN) that propagates labels from unlabelled data through a graph of samples and meta-learns key hyperparameters. Li *et al.* propose a Learning to Self-Train (LST) approach [71] that is based on self-training and meta-learns a soft weighting network to control the influence of pseudo labelled samples. and reduces label-noise during training. The proposed approach Fewmatch (Section 6.2) to SS-FSL differs from the surveyed methods by making use of CR (*c.f.* pseudo-labelling), and therefore alleviates the error propagation problem that is common when pseudo labelling is employed [77]. Furthermore, previous SS-FSL prototypical work [183] only updates prototypes by averaging class representation. Alternatively, Fewmatch iteratively and explicitly refines prototypes using both average feature representation and implicit CR refinement. Experiments show that this enables more flexible feature adaptation to novel tasks and respectively obtains more accurate class prototypes. Furthermore, Fewmatch takes advantage of Imprinted weights [69] in order to provide per class prototypes. One of the main advantages of this prototype-driven method is that it does not require the standard, restrictive episode training strategy that typically frames training as a sequence of artificially designed few-shot learning tasks with consistent number of categories and labelled samples, and imposes the same set-up at test time. This in theory affords us greater flexibility with the learning problem definition, allows for consideration of more practical problem setups, and for easier combination with techniques from other fields such as integration of auxiliary losses.

2.3 Preliminaries of Deep Learning for Object Recognition

This section revisits the vanilla deep model training object function for object recognition and provides the essential mathematical formulations which lay the foundation of approaches provided in this thesis.

For supervised model learning, we assume n labelled training samples $\mathcal{D} = \{(\boldsymbol{I}_i, y_i)\}_i^n$. Each sample belongs to one of C classes $y_i \in \mathcal{Y} = [1, 2, \cdots, C]$, with the ground-truth label typically represented as a one-hot vector. The objective is to learn a classification deep CNN model generalisable to unseen test data through a cost-effective training process.

We begin with reviewing the vanilla deep model training method (Figure 1.5(a)) before elaborating the proposed approaches in this thesis.

2.3.1 Vanilla Deep Model Training

For training a classification deep model, the softmax cross-entropy loss function is usually adopted. Specifically, we predict the posterior probability of a labelled sample $\boldsymbol{I}$ over any class c via the softmax criterion:

$$p(c|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}, \;\; z_j = \boldsymbol{W}_j^\top \boldsymbol{x}, \;\; c \in \mathcal{Y} \tag{2.1}$$

where $\boldsymbol{x}$ refers to the embedded feature vector of $\boldsymbol{I}$, $\boldsymbol{W}_j$ the j-th class prediction function parameter, and $\boldsymbol{\theta}$ the neural network model parameters.

In some cases, we wish that the posterior prediction is not sharp for better transfer the knowledge, thus a *softened* softmax operation is formulated as:

$$\tilde{p}(c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\exp(z_c/T)}{\sum_{j=1}^{C}\exp(z_j/T)}, \ z_j = \boldsymbol{W}_j^{\top}\boldsymbol{x}, \ c \in \mathcal{Y} \tag{2.2}$$

where the tempreture parameters T controls the softening degree.

To train a multi-class classification model, we typically adopt the Cross-Entropy (CE) measurement between the predicted and ground-truth label distributions on a labelled sample $\boldsymbol{x}$ (in a mini-batch) as the objective loss function:

$$\mathcal{L}_{\text{ce}} = -\sum_{c=1}^{C}\delta_{c,y}\log\Big(p(c|\boldsymbol{x},\boldsymbol{\theta})\Big) \tag{2.3}$$

where $\delta_{c,y}$ is Dirac delta which returns 1 if c is the ground-truth label, and 0 otherwise.

Discussion. For a model subject to the vanilla training (Figure 1.5(a)), the cross-entropy loss is utilised to supervise the model parameters (e.g. by the stochastic gradient descent algorithm) iteratively in a *one-stage* procedure. We name this vanilla strategy.

2.3.2 Kullback Leibler (KL) Divergence

In mathematical statistics, the Kullback–Leibler divergence[184] is a measure of the difference between two probability distribution.

Let $p = \{p_1,...,p_j,...,p_C\}$ and $q = \{q_1,...,q_j,..,q_C\}$ are two probability distributions. That is, both p_j and q_j sum up to 1, and $p_j > 0$ and $q_j > 0$ for any $j \in \{1,...,C\}$. Specifically, the Kullback-Leibler (KL) divergence of p from q, denoted as $L_{kl}(p||q)$, which measures the information lost when q is used to approximate p, is formulated as:

$$L_{\text{kl}}(p||q) = \sum_{j=1}^{C} p(x)\log\frac{p(x)}{q(x)}. \tag{2.4}$$

2.4 Summary

The above sections have investigated and discussed most related work to this thesis. In particular, the advanced techniques in fine-grained person re-id and coarse-grained image classification are analysed. Besides, this chapter also discusses several aspects with respect to knowledge transfer, including domain adaptation, knowledge distillation and few shot classification. Despite the promising results achieved by current methods, there are still plenty of weaknesses to overcome and large potential space to improve. Firstly, the inaccurate and multi-scale person bounding box challenge of person re-id and search are discussed. Besides, the lacking of the local constrain in existing domain adaptation methods is analyzed in the context of cross-domain knowledge transfer. Furthermore, the inefficient issue of knowledge distillation and expensive training cost of recently approaches are investigated in the image classification task. At last, this thesis focuses on the limitation of ”single prototype assumption” in the current metric-based learning FSL approaches. In the following chapters, several approaches are proposed to address the above limitations.

Chapter 3

Knowledge Transfer Across Classes in Person Re-identification

Merely quantitative differences, beyond a certain point, pass into qualitative changes.

—— **Karl Marx**

This chapter focuses on cross-class knowledge transfer in person re-identification. As the nature of disjoint person identities between training and testing, cross-class knowledge transfer is in high demand. However, existing person re-id approaches often ignore the influence of background noise and multi-scale of person bounding box to cross-class knowledge transfer, yielding insufficient knowledge transfer. This is more severe in the practice, when the person bounding box must be detected from unconstrained images (*i.e.*, raw video frames), as discussed in Section 1.3.1. To address these limitations, two methods are proposed as the following section: 1) Identity DiscriminativE Attention reinforcement Learning (IDEAL); 2) Cross-Level Semantic Alignment (CLSA).

3.1 Identity DiscriminativE Attention reinforcement Learning (IDEAL)

3.1.1 Model Overview

This section aims to optimise auto-detected person bounding box images for improving cross-class knowledge transfer in re-id matching. To this end, the Identity DiscriminativE Attention

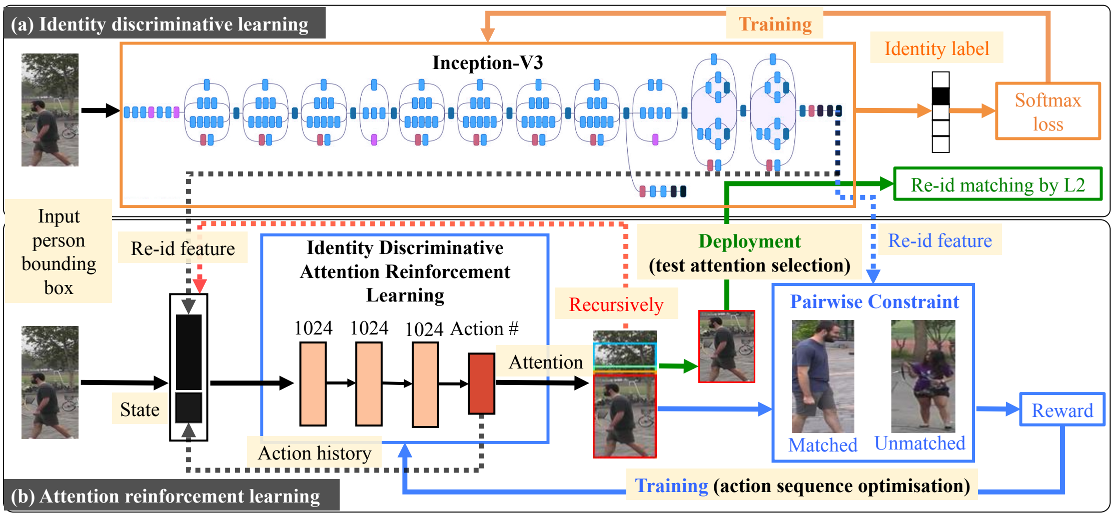
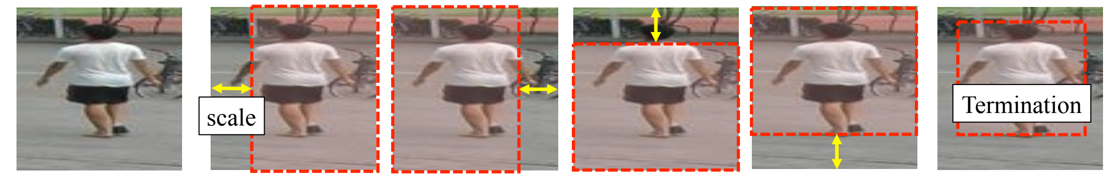

(a) Identity discriminative learning
Training
Inception-V3
Identity label
Softmax loss
Re-id matching by L2
Input person bounding box
Re-id feature
Identity Discriminative Attention Reinforcement Learning
1024 1024 1024 Action #
Deployment (test attention selection)
Re-id feature
Recursively
Pairwise Constraint
Attention
State
Action history
Matched
Unmatched
Reward
Training (action sequence optimisation)
(b) Attention reinforcement learning

Figure 3.1: The IDEAL reinforcement learning attention selection model. (a) An identity discriminative learning branch based on the deep Inception-V3 network optimised by a multi-classification softmax loss (orange arrows). (b) An attention reinforcement learning branch designed as a deep Q-network optimised by re-id class label constraints in the deep feature space from branch (a) (blue arrows). For model deployment, the trained attention branch (b) computes the optimal attention regions for each probe and all the gallery images, extract the deep features from these optimal attention regions in the multi-class re-id branch (a) and perform L2 distance matching (green arrows).

reinforcement Learning (IDEAL) model is formulated. The IDEAL has two sub-networks: **(I)** A multi-class *discrimination network* $\mathcal{D}$ by deep learning from a training set of auto-detected person bounding boxes (Figure 3.1(a)). This part is flexible with many options from existing deep re-id networks and beyond [106, 114, 29, 122]. **(II)** A re-identification *attention network* $\mathcal{A}$ by reinforcement learning recursively a salient sub-region with its deep feature representation from $\mathcal{D}$ that can maximise identity-matching given re-id label constraints (Figure 3.1(b)). Next, we formulate the attention network by reinforcement learning and how this attention network cooperates with the multi-class discrimination network.

3.1.2 Re-ID Attention Selection Formulation

This thesis formulates the re-id attention selection as a reinforcement learning problem [185]. This allows to correlate directly the re-id attention selection process with the learning objective of an "agent" by recursively *rewarding* or punishing the learning process. In essence, the aim of model learning is to achieve an optimal identity attending action policy $a = \pi(\mathbf{s})$ of an agent, i.e. a mapping function, that projects a state observation $\mathbf{s}$ (model input) to an action prediction a. This work exploits the Q-learning technique for learning the proposed IDEAL agent, due to its sample efficiency advantage for a smaller set of actions [186, 187]. Formally, IDEAL aims



(a) Input image (b) Attending actions (Each red dotted box represents the attention window after the action)

Figure 3.2: Identity discriminative attending actions are given by an attending scale variable on four directions (left/right/top/bottom). Termination action means the stop of a recursive attending process.

to learn an optimal state-value function which measures the maximum sum of the current reward (R_t) and all the future rewards ($R_{t+1}, R_{t+2}, \cdots$) discounted by a factor γ at each time step t:

$$Q^*(\mathbf{s},a) = \max_{\pi} \mathbb{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \mid \mathbf{s}_t = \mathbf{s}, a_t = a, \pi\right] \tag{3.1}$$

Once $Q^*(\mathbf{s},a)$ is learned, the optimal policy $\pi^*(\mathbf{s})$ can be directly inferred by selecting the action with the maximum $Q^*(\mathbf{s},a)$ value in model deployment. More specifically, the reinforcement learning agent interacts with each data sample in a sequential episode, which can be considered as a Markov Decision Process (MDP) [188]. Therefore, IDEAL firstly design a specific MDP for re-id discriminative attention selection, as described below.

3.1.3 Markov Decision Process for Re-ID Attention Selection

A MDP is designed for re-id attention selection in auto-detected bounding boxes. In particular, IDEAL considers each input person bounding box image as a dynamic environment. An IDEAL agent interacts with this dynamic environment to locate the optimal re-id attention window. To guide this discriminative learning process, IDEAL further considers a reward that can encourage those attending actions to improve re-id performance and maximise the cumulative future reward in Eq (3.1). As such, actions, states, and rewards are designed as follows.

Actions: An action set $\mathbf{A}$ is defined to facilitate the IDEAL agent to determine the location and size of an "attention window" (Figure 3.2). Specifically, an attending action a is defined by the location shift direction ($a_d \in \{\text{left, right, top, bottom}\}$) and shift scale ($a_e \in \mathbf{E}$). We also introduce a termination action as a search process stopping signal. $\mathbf{A}$ consists of a total of $(4 \times |\mathbf{E}| + 1)$ actions. Formally, let the upper-left and bottom-right corner coordinates of the current attention window and an updated attention window be $[x_1, y_1, x_2, y_2]$ and $[x_1', y_1', x_2', y_2']$ respectively, the action set $\mathbf{A}$ can then be defined as:

$$\mathbf{A} = \{x_1' = x_1 + \alpha\Delta x,\ x_2' = x_2 - \alpha\Delta x,\ y_1' = y_1 + \alpha\Delta y,\ y_2' = y_2 - \alpha\Delta y,\ \mathrm{T}\}, \tag{3.2}$$
$$\text{where}\ \ \alpha \in \mathbf{E},\ \ \Delta x = x_2 - x_1,\ \ \Delta y = y_2 - y_1,\ \ \mathrm{T} = \text{termination}.$$

Computationally, each action except termination in $\mathbf{A}$ modifies the environment by cutting off a horizontal or vertical stripe. IDEAL sets $\mathbf{E} = \{5\%, 10\%, 20\%\}$ by cross-validation in experiments, resulting in total 13 actions. Such a small attention action space with multi-scale changes has three merits: (1) Only a small number of simple actions are contained, which allows more efficient and stable agent training; (2) Fine-grained actions with small attention changes allow the IDEAL agent sufficient freedoms to utilise small localised regions in auto-detected bounding boxes for subtle identity matching. This enables more effective elimination of undesired background clutter whilst retaining identity discriminative information; (3) The termination action enables the agent to be aware of the satisfactory condition met for attention selection and stops further actions when optimised.

States: The state $\mathbf{s}_t$ of our MDP at time t is defined as the concatenation of the feature vector $\mathbf{x}_t \in \mathbb{R}^d$ (with d re-id feature dimension) of current attending window and an action history vector $\mathbf{h}_t \in \mathbb{R}^{|\mathbf{E}| \times n_{\text{step}}}$ (with n_{step} a pre-defined maximal action number per bounding box), i.e. $\mathbf{s}_t = [\mathbf{x}_t, \mathbf{h}_t]$. Specifically, at each time step, we first extract the feature vector $\mathbf{x}_t$ of current attention window by the trained re-id network $\mathcal{D}$. The action history vector $\mathbf{h}_t$ is a binary vector for keeping a track of all past actions, represented by a A-dimensional (13 actions) one-hot vector where the corresponding action bit is encoded as one, all others as zeros.

Rewards: The reward function defines the agent task objective. In this context, this thesis therefore correlates directly the reward function R (Eq (3.1)) of the IDEAL agent's attention behaviour with the re-id matching criterion. Formally, at time step t, suppose the IDEAL agent observes a person image $\mathbf{I}_t$ and then takes an action $a_t = a \in \mathbf{A}$ to attend the image region $\mathbf{I}_t^a$. Given this attention shift from $\mathbf{I}_t$ to $\mathbf{I}_t^a$, its state $\mathbf{s}_t$ changes to $\mathbf{s}_{t+1}$. We need to assess such a state change and signify the agent if this action is encouraged or discouraged by an award or a punishment. To this end, this thesis propose three reward function designs, inspired by pairwise constraint learning principles established in generic information search and person re-id.

Notations From the labelled training data, IDEAL samples two other *reference* images w.r.t. $\mathbf{I}_t$: (1) A *cross-view positive* sample $\mathbf{I}_t^+$ sharing the same identity as $\mathbf{I}_t$ but not the camera view;

(2) A *same-view negative* sample $\mathbf{I}_t^-$ sharing the camera view but not the identity as $\mathbf{I}_t$. iDEAL computes the features of all these images by $\mathcal{D}$, denoted respectively as $\mathbf{x}_t, \mathbf{x}_t^a, \mathbf{x}_t^+$, and $\mathbf{x}_t^-$.

(I) Reward by Relative Comparison The first reward function R_t is based on relative comparison, in spirit of the triplet loss for learning to rank [189]. It is formulated as:

$$R_t = R_{rc}(\mathbf{s}_t, a) = \Big(f_{\text{match}}(\mathbf{x}_t^a, \mathbf{x}_t^-) - f_{\text{match}}(\mathbf{x}_t^a, \mathbf{x}_t^+)\Big) - \Big(f_{\text{match}}(\mathbf{x}_t, \mathbf{x}_t^-) - f_{\text{match}}(\mathbf{x}_t, \mathbf{x}_t^+)\Big) \quad (3.3)$$

where f_{match} defines the re-id matching function. IDEAL uses the Euclidean distance metric given the Inception-V3 deep features. Intuitively, this reward function commits (i) a positive reward if the attended region becomes more-matched to the *cross-view positive* sample whilst less-matched to the *same-view negative* sample, or (ii) a negative reward otherwise. When a is the termination action, i.e. $\mathbf{x}_t^a = \mathbf{x}_t$, the reward value R_{rc} is set to zero. In this way, the IDEAL agent is supervised to attend the regions subject to optimising jointly two tasks: (1) being more discriminative and/or more salient for the target identity in an inter-view sense (cross-view re-id), whilst (2) pulling the target identity further away from other identities in an intra-view sense (discarding likely shared view-specific background clutter and occlusion therefore focusing more on genuine person appearance). Importantly, this multi-task objective design favourably allows appearance saliency learning to intelligently select the most informative parts of certain appearance styles for enabling holistic clothing patten detection and ultimately more discriminative re-id matching (e.g. Figure 1.7(b) and Figure 3.3(b)).

(II) Reward by Absolute Comparison The second reward function considers only the compatibility of a true matching pair, in the spirit of positive verification constraint learning [190]. Formally, this reward is defined as:

$$R_t = R_{ac}(\mathbf{s}_t, a) = \Big(f_{\text{match}}(\mathbf{x}_t, \mathbf{x}_t^+)\Big) - \Big(f_{\text{match}}(\mathbf{x}_t^a, \mathbf{x}_t^+)\Big) \quad (3.4)$$

The intuition is that, the cross-view matching score of two same-identity images depends on how well irrelevant background clutter/occlusion is removed by the current action. That is, a good attending action will increase a cross-view matching score, and vice verse.

(III) Reward by Ranking The third reward function concerns the true match ranking change brought by the agent action, therefore simulating directly the re-id deployment rational [191]. Specifically, IDEAL designs a binary reward function according to whether the rank of true match $\mathbf{x}_t^+$ is improved when $\mathbf{x}_t$ and $\mathbf{x}_t^a$ are used as the probe separately, as:

$$R_t = R_r(\mathbf{s}_t, a) = \begin{cases} +1, & \text{if } \mathrm{Rank}(\mathbf{x}_t^+ | \mathbf{x}_t) > \mathrm{Rank}(\mathbf{x}_t^+ | \mathbf{x}_t^a) \\ -1, & \text{otherwise} \end{cases} \tag{3.5}$$

where $\mathrm{Rank}(\mathbf{x}_t^+|\mathbf{x}_t)$ $(\mathrm{Rank}(\mathbf{x}_t^+|\mathbf{x}_t^a))$ represents the rank of $\mathbf{x}_t^+$ in a gallery against the probe $\mathbf{x}_t$ $(\mathbf{x}_t^a)$. Therefore, Eq (3.5) gives support to those actions of leading to a higher rank for the true match, which is precisely the re-id objective. In the implementation, the gallery was constructed by randomly sampling n_g (e.g. 600) cross-view training samples. The above three reward function choices are evaluted and discussed in the experiments (Sec. 3.1.5).

3.1.4 Model Implementation, Training, and Deployment

Implementation and Training For the multi-class discrimination network $\mathcal{D}$ in the IDEAL model, IDEAL deploys the Inception-V3 network [9], a generic image classification CNN model [9]. It is trained from scratch by a softmax classification loss using person identity labels of the training data. For the re-id attention network $\mathcal{A}$ in the IDEAL model, a neural network of 3 fully-connected layers (each with 1024 neurons) and a prediction layer were designed (Figure 3.1(b)). This implements the state-value function Eq (3.1). For optimising the sequential actions for re-id attention selection, IDEAL utilises sthe ε-greedy learning algorithm [192] during model training: The agent takes (1) a random action from the action set $\mathbf{A}$ with the probability ε, and (2) the best action predicted by the agent with the probability $1-\varepsilon$. IDEAL begin with $\varepsilon = 1$ and gradually decrease it by 0.15 every 1 training epoch until reaching 0.1. The purpose is to balance model exploration and exploitation in the training stage so that local minimum can be avoided. To further reduce the correlations between sequential observations, IDEAL employs the experience replay strategy [192]. In particular, a fixed-sized memory pool $\mathbf{M}$ is created to store the agent's N past training sample (experiences) $e_t = (\mathbf{s}_t, a_t, R_t, \mathbf{s}_{t+1})$ at each time step t, i.e. $\mathbf{M} = \{e_{t-N+1}, \cdots, e_t\}$. At iteration i, a mini-batch of training samples is selected randomly from $\mathbf{M}$ to update the agent parameters θ by the loss function:

$$L_i(\theta_i) = \mathbb{E}_{(\mathbf{s}_t, a_t, R_t, \mathbf{s}_{t+1}) \sim \mathrm{Uniform}(\mathbf{M})} \Big(R_t + \gamma \max_{a_{t+1}} Q(\mathbf{s}_{t+1}, a_{t+1}; \tilde{\theta}_i) - Q(\mathbf{s}_t, a_t; \theta_i) \Big)^2, \tag{3.6}$$

where $\tilde{\theta}_i$ are the parameters of an intermediate model for predicting training-time target values, which are updated as θ_i at every ς iteration, but frozen at other times.

Deployment During model deployment, IDEAL applys the learned attention network $\mathcal{A}$ to all test probe and gallery bounding boxes for extracting their attention window images. The deep features of these attention window images are used for person re-id matching by extracting the 2,048-D output from the last fully-connected layer of the discrimination network $\mathcal{D}$. IDEAL employs the L2 distance as the re-id matching metric.

3.1.5 Experiments

Datasets For evaluation, IDEAL used two large benchmarking re-id datasets generated by automatic person detection: CUHK03 [80], and Market-1501 [43] (details in Table 2.1). CUHK03 also provides an extra version of bounding boxes by human labelling therefore offers a like-to-like comparison between the IDEAL attention selection and human manually cropped images. Example images are shown in (a),(b) and (c) of Figure 1.7.

Evaluation Protocol We adopted the standard CUHK03 1260/100 training/test split [80] and the standard Market-1501 training/test split (750/751) [43] for the single-query evaluation setting. We used the cumulative matching characteristic (CMC) to measure re-id accuracy. For the Market-1501 multi-query setting, we also used the recall measure of truth matches by mean Average Precision (mAP), i.e. computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes.

Table 3.1: Comparing re-id performance. 1$^{\text{st}}$/2$^{\text{nd}}$ best results are shown in red/blue. AD: Automatically Detected;

Dataset	CUHK03(AD) [80]				Market-1501(AD) [43]			
Metric (%)	R1	R5	R10	R20	R1	mAP	R1	mAP
ITML [193]	5.1	17.7	28.3	-	-	-	-	-
LMNN [195]	6.3	18.7	29.0	-	-	-	-	-
KISSME [100]	11.7	33.3	48.0	-	40.5	19.0	-	-
MFA [104]	-	-	-	-	45.7	18.2	-	-
kLFDA [104]	-	-	-	-	51.4	24.4	52.7	27.4
BoW [43]	23.0	42.4	52.4	64.2	34.4	14.1	42.6	19.5
XQDA [42]	46.3	78.9	83.5	93.2	43.8	22.2	54.1	28.4
MLAPG [108]	51.2	83.6	92.1	**96.9**	-	-	-	-
L$_1$-Lap [199]	30.4	-	-	-	-	-	-	-
NFST [111]	53.7	83.1	93.0	94.8	55.4	29.9	68.0	41.9
LSSCDL [201]	51.2	80.8	89.6	-	-	-	-	-
SCSP [202]	-	-	-	-	51.9	26.3	-	-
TMA [194]	-	-	-	-	47.9	22.3	-	-
HL [196]	-	-	-	-	59.5	-	-	-
HER [122]	60.8	87.0	**95.2**	**97.7**	-	-	-	-
FPNN [80]	19.9	-	-	-	-	-	-	-
DCNN+ [113]	44.9	76.0	83.5	93.2	-	-	-	-
EDM [197]	52.0	-	-	-	-	-	-	-
SICI [114]	52.1	84.9	92.4	-	-	-	-	-
SSDAL [198]	-	-	-	-	39.4	19.6	49.0	25.8
S-LSTM [200]	57.3	80.1	88.3	-	-	-	61.6	35.3
eSDC [126]	7.7	21.9	35.0	50.0	33.5	13.5	-	-
CAN [131]	63.1	82.9	88.2	93.3	48.2	24.4	-	-
Gated S-CNN [123]	**68.1**	**88.1**	**94.6**	-	**65.8**	**39.5**	**76.0**	**48.4**
IDEAL	**71.0**	**89.8**	93.0	95.9	**86.7**	**67.5**	**91.3**	**76.2**

Implementation Details We implemented the proposed IDEAL method in the TensorFlow framework [203]. We trained an Inception-V3 [9] multi-class identity discrimination network $\mathcal{D}$ from scratch for each re-id dataset at a learning rate of 0.0002 by using the Adam optimiser

[204]. The final FC layer output feature vector (2,048-D) together with the L2 distance metric is used as our re-id matching model. All person bounding boxes were resized to 299×299 in pixel. We trained the $\mathcal{D}$ by 100,000 iterations. We optimised the IDEAL attention network $\mathcal{A}$ by the Stochastic Gradient Descent algorithm [205] with the learning rate set to 0.00025. We used the relative comparison based reward function (Eq (3.3)) by default. The experience replay memory size $\mathbf{M}$ for reinforcement learning was 100,000. We fixed the discount factor γ to 0.8 (Eq (3.1)). We allowed a maximum of $n_{\text{step}} = 5$ action rounds for each episode in training $\mathcal{A}$. The intermediate regard prediction network was updated every $\varsigma = 100$ iterations. We trained the $\mathcal{A}$ by 10 epochs.

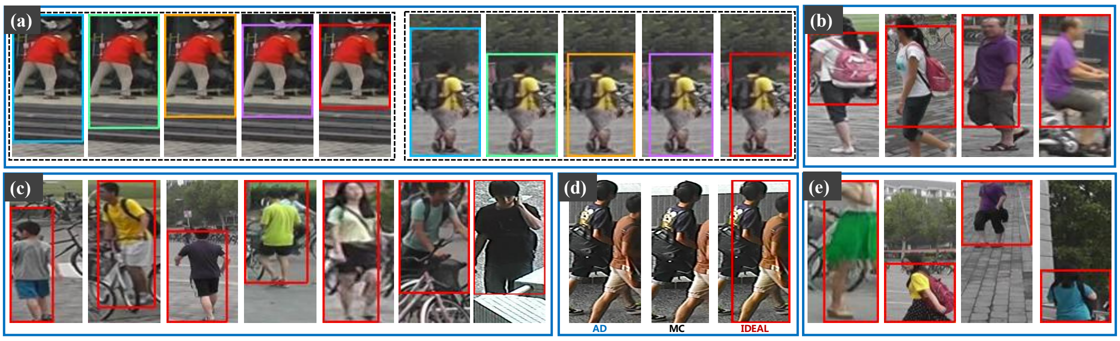


Figure 3.3: Qualitative evaluations of the IDEAL model: **(a)** Two examples of action sequence for attention selection given by action1 (Blue), action2 (Green), action3 (Yellow), action4 (Purple), action5 (Red); **(b)** Two examples of cross-view IDEAL selection for re-id; **(c)** Seven examples of IDEAL selection given by 5, 3, 5, 5, 4, 2, and 2 action steps respectively; **(d)** A failure case when the original auto-detected (AD) bounding box contains two people, manually cropped (MC) gives a more accurate box whilst IDEAL attention selection fails to reduce the distraction; **(e)** Four examples of IDEAL selection on the Market-1501 "distractors" with significantly poorer auto-detected bounding boxes when IDEAL shows greater effects.

Comparisons to the State-of-the-Arts We compared the IDEAL model against 24 different contemporary and the state-of-the-art re-id methods (Table 3.1). It is evident that IDEAL achieves the best re-id performance, outperforming the strongest competitor Gated S-CNN [123] by 2.9% (71.0-68.1) and 20.9% (86.7-65.8) in Rank-1 on CUHK03 and Market-1501 respectively. This demonstrates a clear positive effect of IDEAL's attention selection on person re-id performance by filtering out bounding box misalignment and random background clutter in auto-detected person images. To give more insight and visualise both the effect of IDEAL and also failure cases, qualitative examples are shown in Figure 3.3.

Evaluations on Attention Selection We further compared in more details the IDEAL model

Table 3.2: Comparing attention selection methods. SQ: Single Query; MQ: Multi-Query.

Dataset	CUHK03 [80]				Market-1501 [43]			
Metric (%)	R1	R5	R10	R20	R1(SQ)	mAP(SQ)	R1(MQ)	mAP(MQ)
eSDC [126]	7.7	21.9	35.0	50.0	33.5	13.5	-	-
CAN [131]	63.1	82.9	88.2	93.3	48.2	24.4	-	-
Gated S-CNN [123]	68.1	88.1	**94.6**	-	65.8	39.5	76.0	48.4
No Attention	67.5	88.2	92.6	95.7	84.5	64.8	89.4	72.5
Random Attention	54.1	79.2	85.9	90.4	80.3	54.6	85.1	66.7
Centre Attention (95%)	66.1	86.7	91.1	94.9	84.1	64.2	88.6	69.4
Centre Attention (90%)	64.1	85.3	90.3	93.5	82.7	60.3	87.5	65.3
Centre Attention (80%)	51.9	76.0	83.0	89.0	74.7	48.5	83.4	57.6
Centre Attention (70%)	35.2	62.3	73.2	81.7	63.8	39.0	72.3	43.5
Centre Attention (50%)	16.7	38.8	49.5	62.5	39.9	18.5	46.3	23.9
IDEAL(Ranking)	70.3	89.1	92.7	95.4	86.2	66.3	90.8	74.3
IDEAL(AC)	69.1	88.4	92.1	95.0	85.3	65.5	87.5	72.3
IDEAL(RC)	**71.0**	**89.8**	93.0	**95.9**	**86.7**	**67.5**	**91.3**	**76.2**

against three state-of-the-art saliency/attention based re-id models (eSDC [126], CAN [131], Gated S-CNN [123]), and two baseline attention methods (Random, Centre) using the Inception-V3 re-id model (Table 3.2). For *Random Attention*, we attended randomly person bounding boxes by a ratio (%) randomly selected from $\{95, 90, 80, 70, 50\}$. We repeated 10 times and reported the mean results. For *Centre Attention*, we attended all person bounding boxes at centre by one of the same 5 ratios above. It is evident that the IDEAL (Relative Comparison) model is the best. The inferior re-id performance of eSDC, CAN and Gated S-CNN is due to their strong assumption on accurate bounding boxes. Both Random and Centre Attention methods do not work either with even poorer re-id accuracy than that with "No Attention" selection. This demonstrates that optimal attention selection given by IDEAL is non-trivial. Among the three attention reward functions, *Absolute Comparison* is the weakest, likely due to the lack of reference comparison against false matches, i.e. no population-wise matching context in attention learning. *Ranking* fares better, as it considers reference comparisons. The extra advantage of *Relative Comparison* is due to the *same-view negative* comparison in Eq (3.3). This provides a more reliable background clutter detection since same-view images are more likely to share similar background patterns.

Auto-Detection+IDEAL vs. Manually Cropped Table 3.3 shows that auto-detection+IDEAL can perform similarly to that of *manually cropped* images in CUHK03 test[1], e.g. 71.0% vs. 71.9% for Rank-1 score. This shows the potential of IDEAL in eliminating expensive manual labelling of bounding boxes and for scaling up re-id to large data deployment.

Table 3.3: Auto-detection+IDEAL vs. manually cropped re-id on CUHK03.

Metric (%)	R1	R5	R10	R20
Auto-Detected+**IDEAL**	71.0	89.8	93.0	95.9
Manually Cropped	**71.9**	**90.4**	**94.5**	**97.1**

Effect of Action Design We examined three action designs with different size scales. Table 3.4 shows that the most fine-grained design $\{5\%, 10\%, 20\%\}$ is the best. This suggests that the re-id by appearance is subtle and small regions make a difference in discriminative matching.

Table 3.4: Attention action design evaluation. SQ: Single Query; MQ: Multi-Query.

Dataset	CUHK03 [80]				Market-1501 [43]			
Metric (%)	R1	R5	R10	R20	R1(SQ)	mAP(SQ)	R1 (MQ)	mAP(MQ)
{5%, 10%, 20%}	**71.0**	**89.8**	**93.0**	**95.9**	**86.7**	**67.5**	**91.3**	**76.2**
{10%, 20%, 30%}	68.3	88.1	91.8	95.0	86.2	66.8	90.5	73.4
{10%, 20%, 50%}	67.6	87.5	91.4	93.9	85.3	65.6	88.8	72.1

3.2 Person search by multi-scale matching

3.2.1 Cross-Level Semantic Alignment for Person Search

This chapter further considers more realistic problem: person search, and focus on the negative influence of multi-scale person bounding box to the cross-class knowledge transfer. We want to establish a person search system capable of automatically detecting and matching persons in unconstrained scenes with any probe person. With the arbitrary distances between people and cameras in public space, person images are inherently captured at varying scales and resolutions. This multi-scale matching challenge will influence the cross-class knowledge transfer To overcome this problem, we formulate a Cross-Level Semantic Alignment (CLSA) deep learning approach. An overview of the CLSA is illustrated in Figure 3.4. The CLSA contains two components: (1) Person detection which locates all person instances in the gallery scene images for facilitating the subsequent identity matching. (2) Person re-identification which matches the

[1] The Market-1501 dataset provides no manually cropped person bounding boxes.

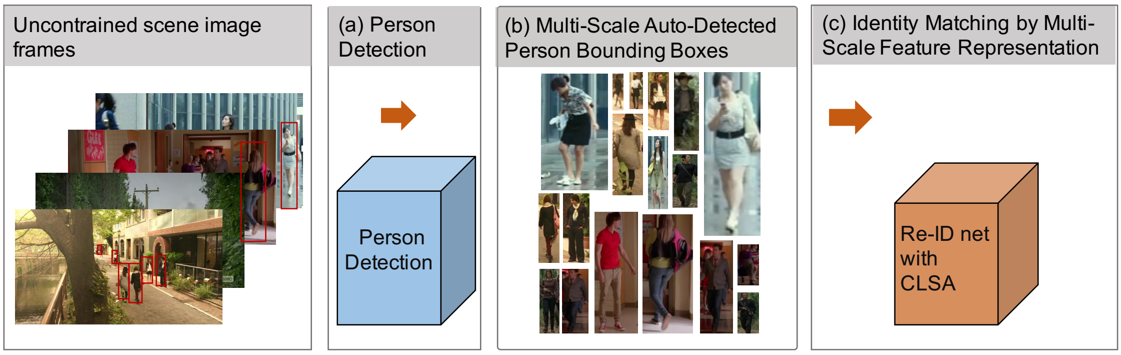

Uncontrained scene image frames
(a) Person Detection
Person Detection
(b) Multi-Scale Auto-Detected Person Bounding Boxes
(c) Identity Matching by Multi-Scale Feature Representation
Re-ID net with CLSA

Figure 3.4: An overview of the proposed multi-scale learning person search framework. (a) Person detection for cropping people from the whole scene images at (b) varying scales (resolutions). (c) Person identity matching is then conducted by a re-id model.

probe image against a large number of arbitrary scale gallery person bounding boxes (the key component of CLSA). We provide the component details below.

Person Detection

As a pre-processing step, person detection is important in order to achieve accurate search [28, 37]. We adopt the Faster-RCNN model [206] as the CLSA detection component, due to its strong capability of detecting varying sized objects in unconstrained scenes. To further enhance person detection performance and efficiency, we introduce a number of design improvement on the original model. **(1)** Instead of using the conventional RoI (Region of Interest) pooling layer, we crop and resize the region feature maps to 14×14 in pixel, and further max-pool them to 7×7 for gaining better efficiency [207]. **(2)** After pre-training the backbone ResNet-50 net on ImageNet-1K, we fix the 1^{st} building-block (the 1^{st} 4 layers) in fine-tuning on the target person search data. This allows to preserve the shared low-level features learned from larger sized source data whilst simultaneously adapting the model to target data. **(3)** We keep and exploit all sized proposals for reducing the mis-detection rate at extreme scales in uncontrolled scenes before the Non-Maximum Suppression (NMS) operation. In deployment, we consider all detection boxes scored above 0.5, rather than extracting a fixed number of boxes from each scene image [37]. This is because the gallery scene images may contain varying (unknown in priori) number of people.

Multi-Scale Matching by Cross-Level Semantic Alignment

Given auto-detected person bounding boxes at arbitrary scales from the gallery scene images, we aim to build a person identity search model robust for multi-scale matching. To this end, we explore the seminal image/feature pyramid concept [208, 209, 84, 83]. Our motivation is that a single-scale feature representation blurs salient and discriminative information at different scales useful in person identity matching; And a pyramid representation allows to be "scale-invariant" (more "scale insensitive") in the sense that a scale change in matching images is counteracted by a scale shift within the feature pyramid.

Build-In Feature Pyramid We investigate the multi-scale feature representation learning in deep Convolutional Neural Network (CNN) to exploit the built-in feature pyramid structure formed on a single input image scale. Although CNN features have shown to be more robust to variance in image scale, pyramids are still effective in seeking more accurate detection and recognition results [210].

For the CNN architecture, we adopt the state-of-the-art ResNet-50 [5] as the backbone network (Figure 3.5) of the identity matching component. In this study, we particularly leverage the feature pyramid hierarchy with low-to-high levels of semantics from bottom to top layers, automatically established in model learning optimisation [211]. Given the block-wise net structure in ResNet-50, we build a computationally efficient K-levels feature pyramid using the last conv layer of top-K ($K = 3$ in our experiments) blocks. The deepest layer of each block is supposed to have the most semantic features.

Nonetheless, it is not straightforward to exploit the ResNet-50 feature hierarchy. This is because the build-in pyramid has large semantic gaps across levels due to the distinct depths of layers. The features from lower layers are less discriminative for person matching therefore likely hurt the overall representational capacity if applied jointly with those from higher layers.

Cross-Level Semantic Alignment To address the aforementioned problems, we improve the in-network feature pyramid by introducing a Cross-Level Semantic Alignment (CLSA) learning mechanism. The aim is to achieve a feature pyramid with all levels encoding the desired high-level person identity semantics. Formally, to train our person identity matching model, we adopt the softmax Cross-Entropy (CE) loss function to optimise an identity classification task. The CE loss on a training person bounding box $(\boldsymbol{I}, y)$ according Eq (2.1) and (2.3) is computed as:

$$\mathcal{L}_{\text{ce}} = -\log\Big(\frac{\exp(\boldsymbol{W}_y^\top \boldsymbol{x})}{\sum_{i=1}^{|\mathcal{Y}|} \exp(\boldsymbol{W}_i^\top \boldsymbol{x})}\Big) \tag{3.7}$$

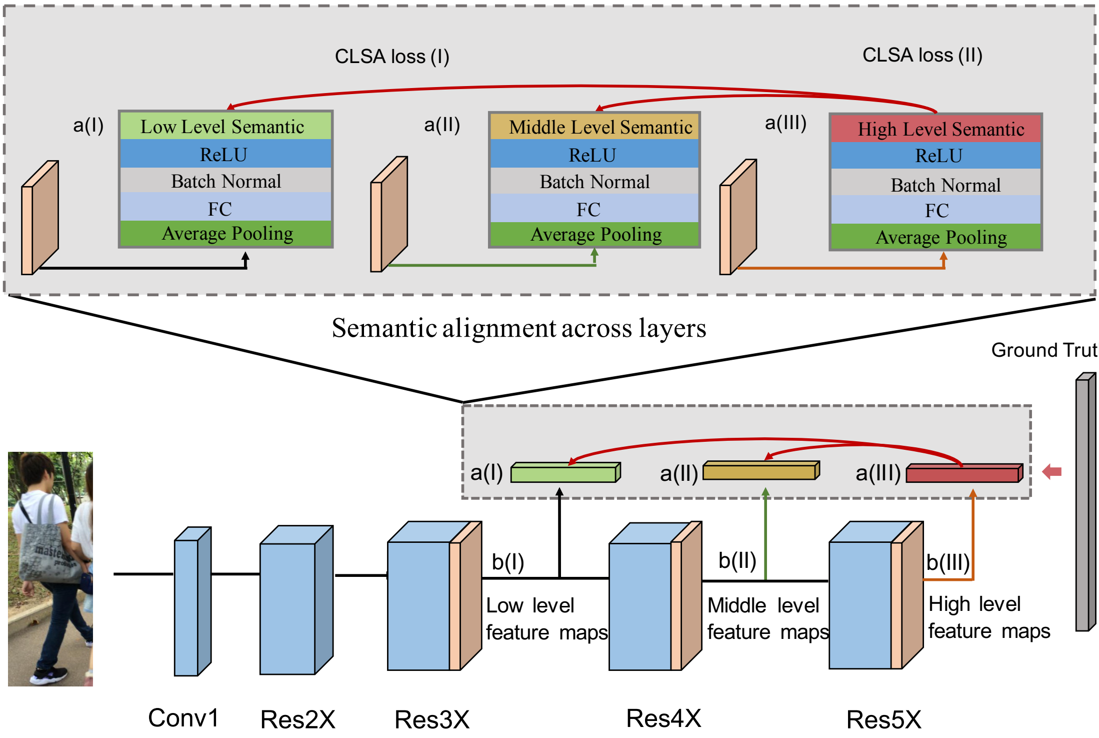


Figure 3.5: An overview of the proposed Cross-Level Semantic Alignment (CLSA) approach in a ResNet-50 based implementation.

where $\boldsymbol{x}$ specifies the feature vector of $\boldsymbol{I}$ by the last layer, $\mathcal{Y}$ the training identity class space, and $\boldsymbol{W}_y$ the y-th ($y \in \mathcal{Y}$) class prediction function parameters.

In our case, $\boldsymbol{x}$ is the top pyramid level, also denoted as $\boldsymbol{x}^K$. For anyone of the top-K ResNet blocks, we obtain $\boldsymbol{x}$ by applying an average pooling layer and a FC layer on the output feature maps (Figure 3.5 (b)). Consider the different feature scale distributions across layers [212], we further normalise $\boldsymbol{x}$ by batch normalisation and ReLU non-linearity. In this way, we compute the feature representations for all K pyramid layers $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^K\}$.

Recall that we aim to render all levels of feature representations identity semantic. To this end, we first project each of these features $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^K\}$ by a FC layer into the identity semantic space with the same dimension as $\mathcal{Y}$. The resulted semantic class probability vectors are denoted as $\{\boldsymbol{p}^1, \cdots, \boldsymbol{p}^K\}$ with $\boldsymbol{p}^k = [p_1^k, \cdots, p_{|\mathcal{Y}|}^k]$, $k \in \{1, \cdots, K\}$. To transfer the strongest semantics from the top (K-th) pyramid level to a lower (s-th) level, we introduce a Kullback-Leibler divergence (Eq (2.4)) based Cross-Level Semantic Alignment (CLSA) loss formulation inspired by knowledge distillation [25]:

$$\mathcal{L}_{\text{clsa}}(s) = \sum_{j=1}^{|\mathcal{Y}|} \tilde{p}_j^K \log \frac{\tilde{p}_j^K}{\tilde{p}_j^s}. \tag{3.8}$$

where $\tilde{p}_j^k$ is a *softened* per-class prediction semantic score obtained by

$$\tilde{p}_j^k = \frac{\exp(p_j^k/T)}{\sum_{j=1}^{|\mathcal{Y}|} \exp(p_j^k/T)}, \tag{3.9}$$

where the temperature parameter T controls the softening degree (higher values meaning more softened predictions). We set T=3 following the suggestion in [25]. To enable end-to-end deep learning, we add this CLSA loss on top of the conventional CE loss (Eq (3.7)):

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + T^2 \sum_{s=1}^{K-1} \mathcal{L}_{\text{clsa}}(s) \tag{3.10}$$

where T^2 serves as a weighting parameter between the two loss terms.

Identity Matching by CLSA Feature Pyramid In deployment, we first compute a CLSA feature pyramid by forward propagating any given person bounding box image. We then concatenate the feature vectors of all pyramid levels as the final representation for person re-id matching.

Remarks The CLSA is similar in spirit to a few person re-id matching methods [137, 139]. However, these methods adopt the image pyramid scheme, in contrast to the CLSA leveraging the in-network feature pyramid on a single image scale therefore more efficient. The FPN model [210] also exploits the build-in pyramid. The CLSA differs from FPN in a number of fundamental ways: (1) FPN focuses on object detection and segmentation, whilst CLSA aims to address

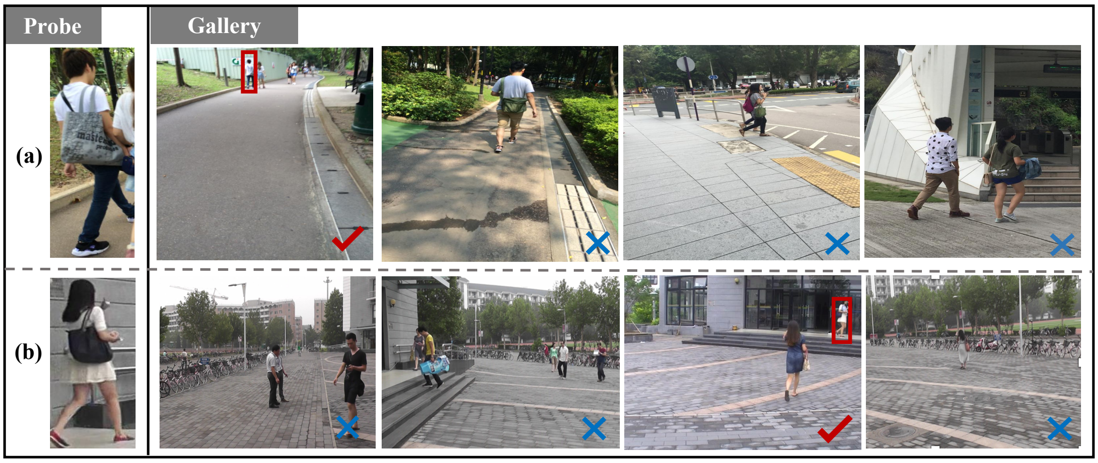


Figure 3.6: Example probe person and unconstrained scene images on (a) CUHK-SYSU [28] and (b) PRW [37]. Green bounding box: the ground truth probe person in the scene. ✓: Contain the probe person. ✗: Not contain the probe person.

fine-grained identity recognition and matching. (2) FPN additionally performs feature map unsampling hence less efficient than CLSA. (3) CLSA performs semantic alignment and transfer in the low-dimensional class space, in comparison to more expensive FPN's feature alignment. We will evaluate and compare these multi-scale learning methods against CLSA in our experiments (Table 3.8).

3.2.2 Experiments

Datasets To evaluate the CLSA, we selected two person search benchmarks: CUHK-SYSU [28] and PRW [37]. We adopted the standard evaluation setting as summarised in Table 3.5. In particular, the CUHK-SYSU dataset contains 18,184 scene images, 8,432 labelled person IDs, and 96,143 annotated person bounding boxes. Each probe person appears in two or more scene gallery images captured from different locations. The training set has 11,206 images and 5,532 probe persons. Within the testing set, the probe set includes 2,900 person bounding boxes and the gallery contains a total of 6,978 whole scene images. The PRW dataset provides a total of 11,816 video frames and 43,110 person bounding boxes. The training set has 482 different IDs from 5,704 frames. The testing set contains 2,057 probe people along with a gallery of 6,112 scene images. In terms of bounding box scale, CUHK-SYSU and PRW range from 37×13 to 793×297, and 58×21 to 777×574, respectively. This shows the two person search datasets present the intrinsic multi-scale challenge. Example images are shown in Figure 3.6.

Performance Metrics For person detection, a person box is considered as correct if overlapping with the ground truth over 50% [28, 37]. For person identity matching or re-id, we adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). The CMC is computed on each individual rank k as the probe cumulative percentage of truth matches appearing at ranks $\leq k$. The mAP measures the recall of multiple truth matches, computed by first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes [43].

Table 3.5: Evaluation setting, data statistics, and person bounding box scale of the CUHK-SYSU and PRW benchmarks. Bbox: Bounding box.

Dataset	Images	Bboxes	IDs	**Bbox Scale**	ID Split		Bbox Split	
					Train	Test	Train	Test
CUHK-SYSU[28]	18,184	96,143	8,432	$37\times13\sim793\times297$	5,532	2,900	55,272	40,871
PRW[37]	11,816	43,110	932	$58\times21\sim777\times574$	482	450	18,048	25,062

Implementation Details We adopted the Pytorch framework [213] to conduct all the following experiments. For training the person detector component, we adopted the SGD algorithm with the momentum set to 0.9, the weight decay to 0.0001, the iteration to 110,000, and the batch size to 256. We initialised the learning rate at 0.001, with a decay factor of 10 at every 30,000 iterations. For training the identity matching component, we used both annotated and detected (over 50% Intersection over Union (IoU) with the annotated and sharing the identity labels) boxes as [37]. We set the momentum to 0.9, the weight decay to 0.00001, the batch size to 64, and the epoch to 100. The initial learning rate was set at 0.01, and decayed by 10 at every 40 epochs. All person bounding boxes were resized to 256 × 128 pixels. To construct the in-network feature pyramid, we utilised the top 3 (Res3x, Res4x, Res5x) blocks in our final model implementation, i.e. $K = 3$ in Eq (3.10). We also evaluated other pyramid constructing ways in the component analysis (Table 3.2.2).

Comparisons to State-Of-The-Art Person Search Methods

We compared the proposed CLSA method with two groups of existing person search approaches: (1) Three most recent state-of-the-art methods (NPSM [38], OIM [28], CWS [37]); and (2) Five popular person detectors (DPM [119], ACF [214], CCF [215], LDCF [216], and R-CNN [217]) with hand-crafted (BoW [43], LOMO [42], DenseSIFT-ColorHist (DSIFT) [218]) or deep learn-

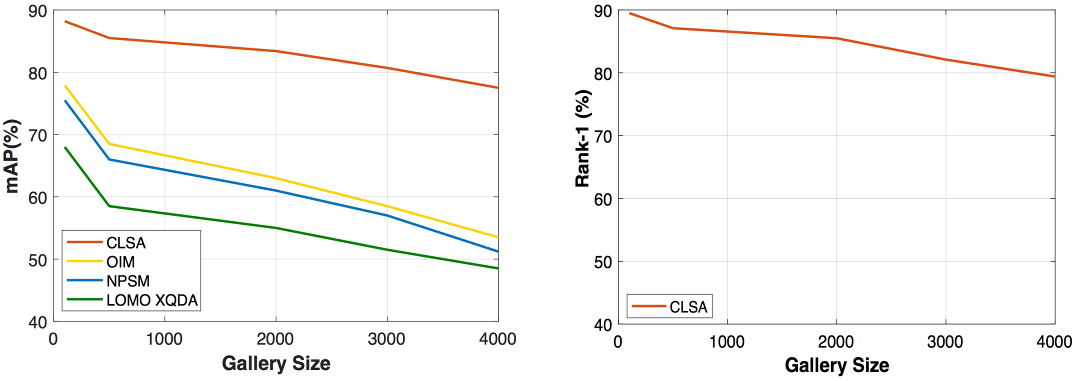


Figure 3.7: Model scalability evaluation over different gallery search sizes on CUHK-SYSU.

ing (IDNet [28]) features based re-id metric learning methods (KISSME [219], XQDA [42]).

Evaluation on CUHK-SYSU Table 3.6 reports the person search performance on CUHK-SYSU with the standard gallery size of 100 scene images. It is clear that the CLSA significantly outperforms all other competitors. For instance, the CLSA surpasses the top-2 alternative models NPSM and OIM (both are end-to-end deep learning models) by 7.3% (88.5-81.2) and 9.8% (88.5-78.7) in Rank-1, 9.3% (87.2-77.9) and 11.7% (87.2-75.5) in mAP, respectively. The performance margin of CLSA against other non-deep-learning methods is even larger, due to that these models rely on less discriminative hand-crafted features without the modelling advantage of jointly learning stronger representation and matching metric model. This shows the overall performance superiority of the CLSA over current state-of-the-art methods, thanks to the joint contributions of improved person detection model (see more details below) and the proposed multi-scale deep feature representation learning mechanism.

To evaluate the model efficiency, we conducted a person search test among 100 gallery images on CUHK-SYSU. We deployed a desktop with a Nvidia Titan X GPU. Applying CLSA, OIM, and NPSM takes 1.2, 0.8, and 120 seconds, respectively. This indicates that the performance advantages of our CLSA do not sacrifice the model efficiency.

To test the model performance scalability, we further evaluated top-3 methods under varying gallery sizes in the range from 100 to 4,000 (the whole test gallery set). We observed in Figure 3.7 that all methods degrade the performance given larger gallery search pools. When increasing the gallery size from 100 to 4,000, the mAP performance of NPSM drops from 77.9% to 53.0%, i.e. -24.9% degradation (no reported Rank-1 results). In comparison, the CLSA is more robust against the gallery size, with mAP/Rank-1 drop at -9.7% (77.5-87.2) and -9.1% (79.4-88.5). This is primarily because more distracting people are involved in the identity matching

Table 3.6: Evaluation on CUHK-SYSU. Gallery size: 100 scene images. The best and second-best results are in red and blue.

Method	Rank-1 (%)	mAP (%)
ACF[214]+DSIFT[218]+Euclidean	25.9	21.7
ACF[214]+DSIFT[218]+KISSME[219]	38.1	32.3
ACF[214]+LOMO[42]+XQDA[42]	63.1	55.5
CCF[215] +DSIFT[218]+Euclidean	11.7	11.3
CCF[215]+DSIFT[218]+KISSME[219]	13.9	13.4
CCF[215]+LOMO[42]+XQDA[42]	46.4	41.2
CCF[215]+IDNet[28]	57.1	50.9
CNN[206]+DSIFT[218]+Euclidean	39.4	34.5
CNN[206]+DSIFT[218]+KISSME[219]	53.6	47.8
CNN[206]+LOMO[42]+XQDA[42]	74.1	68.9
CNN[206]+IDNet[28]	74.8	68.6
OIM[28]	78.7	75.5
NPSM[38]	**81.2**	**77.9**
CLSA	**88.5**	**87.2**

process, presenting more challenging tasks. Importantly, the performance gain of CLSA over other competitors becomes even higher at larger search scales, desirable in real-world applications. This indicates the superior deployment scalability and robustness of CLSA over existing methods in tackling a large scale person search problem, further showing the importance of solving the previously ignored multi-scale matching challenge given auto-detected noisy bounding boxes in person search.

Evaluation on PRW We further evaluated the CLSA against 11 existing competitors on the PRW dataset under the benchmarking setting with 11,816 gallery scene images. Overall, we observed similar performance comparisons with the state-of-the-art methods as on CUHK-SYSU. In particular, the CLSA is still the best person search performer with significant accuracy margins over other alternative methods, surpassing the second-best model NPSM by 11.9% (65.0-53.1) and 14.5% (38.7-24.2) in Rank-1 and mAP, respectively. This consistently suggests the model design advantages of CLSA over existing person search methods in a different video surveillance

Table 3.7: Evaluation on PRW. The best and second-best results are in red and blue.

Method	Rank-1 (%)	mAP (%)
ACF-Alex[214]+LOMO[42]+XQDA[42]	30.6	10.3
ACF-Alex[214]+IDE$_{det}$[37]	43.6	17.5
ACF-Alex[214]+IDE$_{det}$[37] +CWS [37]	45.2	17.8
DPM-Alex[119]+LOMO[42]+XQDA[42]	34.1	13.0
DPM-Alex[119]+IDE$_{det}$[37]	47.4	20.3
DPM-Alex[119]+IDE$_{det}$[37]+CWS[37]	48.3	20.5
LDCF[216]+LOMO[42]+XQDA[42]	31.1	11.0
LDCF[216]+IDE$_{det}$[37]	44.6	18.3
LDCF[216]+IDE$_{det}$[37] +CWS[37]	45.5	18.3
OIM[28]	49.9	21.3
NPSM[38]	**53.1**	**24.2**
CLSA	**65.0**	**38.7**

scenario.

Comparisons to Alternative Multi-Scale Learning Methods

Apart from existing person search methods, we further evaluated the effectiveness of CLSA by comparing with the in-network feature pyramid (baseline) and four state-of-the-art multi-scale deep learning approaches including DeepMu [220], MST [221], DPFL [137], and FPN [210] on the CUHK-SYSU benchmark. We used the standard 100 sized gallery setting in this test. For all compared methods, we utilised the same person detection model and the same backbone identity matching network (except DeepMu [220] that exploits a specially proposed CNN architecture) as the CLSA for fair comparison.

Table 3.8 shows that the proposed CLSA is more effective than other multi-scale learning algorithms in person search. In particular, we have these observations: **(1)** The in-network feature pyramid decreases the overall performance as compared to using the standard ResNet-50 features (no pyramid) by a margin of 1.4% (82.5-81.1) in Rank-1 and 1.4% (81.6-80.2) in mAP. This verifies our hypothesis that directly applying the CNN feature hierarchy may harm the model performance due to the intrinsic semantic discrepancy across different pyramid levels. **(2)** CLSA improves the baseline in-network feature pyramid by a gain of 7.4% (88.5-81.1) in Rank-1 and

Table 3.8: Evaluating different multi-scale deep learning methods on CUHK-SYSU in the standard 100 sized gallery setting. FLOPs: FLoating point OPerations.

Method	Rank-1 (%)	mAP (%)	FLOPs ($\times 10^9$)
ResNet-50	82.5	81.6	**2.678**
In-Network Pyramid	81.1	80.2	**2.678**
DeepMu [220]	78.3	75.8	-
MST [221]	82.7	81.9	8.034
DPFL [137]	84.7	83.8	5.400
FPN [210]	85.5	85.0	4.519
CLSA	**88.5**	**87.2**	2.680

7.0% (87.2-80.2) in mAP. This indicates the exact effectiveness of the proposed cross-level semantic alignment mechanism in enhancing the person identity matching capability of the CNN feature representation in an end-to-end learning manner. **(3)** Three ResNet-50 based competitors all bring about person search performance improvement although less significant than the CLSA. This collectively suggests the importance of addressing the multi-scale matching problem in person search. **(4)** For model computational efficiency in FLOPs (FLoating point OPerations) per bounding box, CLSA has the least (a marginal) cost increase compared to other state-of-the-art multi-scale learning methods. This shows the superior cost-effectiveness of CLSA over alternative methods in addition to its accuracy advantages.

Further Analysis and Discussions

Effect of Person Detection We analysed the effect of person detection on the person search performance using the CUHK-SYSU benchmark. We started with the three customised components of Faster-RCNN (Sec 3.2.1). Table 3.9 shows that: **(1)** The region proposal resizing and max-pool operation does not hurt the model performance. In effect, this is a replacement of ROI pooling. In the context of an average pooling to 1×1 feature map followed, such a design remains the capability of detecting small objects therefore imposing no negative effect. **(2)** Freezing the first block's parameters in fine-tuning detector helps due to the commonality of source and target domain data in low-level feature patterns. **(3)** Using all sized proposals improves the result. It is worthy noting this does not reduce the model efficiency, because only top 256 boxes per image are remained after the Non-Maximum Suppression operation, similar to the conventional

Table 3.9: Detection model component analysis on CUHK-SYSU.

	Full		No resize&max-pool		Not fix 1st block		Not all sized proposals	
Metric (%)	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CLSA	**88.5**	87.2	88.3	**87.3**	87.7	86.8	87.9	86.9

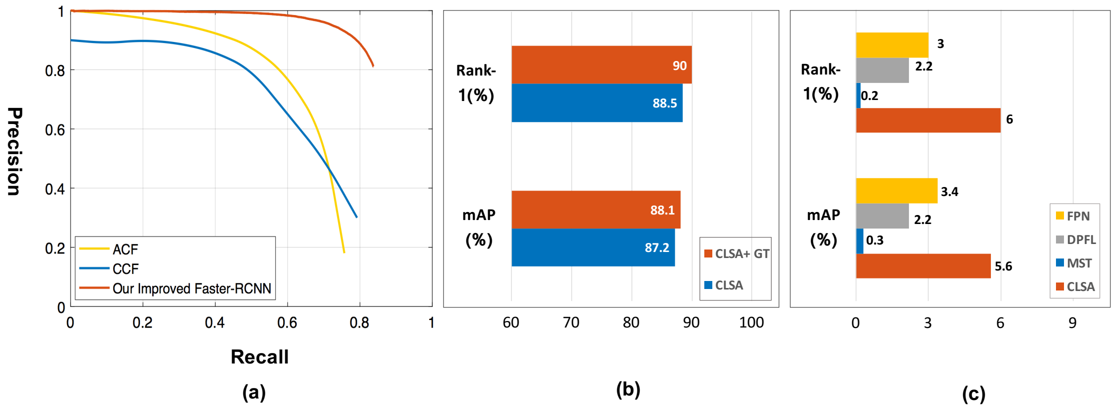


Figure 3.8: Evaluation of person detection on CUHK-SYSU in the standard 100 sized gallery setting. **(a)** Person detection precision-recall performance. **(b)** The person search performance of the CLSA based on auto-detected *or* ground-truth person bounding box images. **(c)** Person detection *versus* multi-scale learning on the effect of person search performance.

case of selecting larger proposals. There are an average of 6.04 bounding boxes per image on CUHK-SYSU.

We then evaluated the holistic person detection performance with comparison to other two detection models (ACF [214] and CCF [215]). For person detection, it is shown in Figure 3.8 (a) that the precision performance of both ACF and CCF drops quickly when increasing the recall rate, whilst our improved Faster-RCNN remains more stable. This shows the effectiveness of deep learning detectors along additional model improvement from our CLSA. This is consistent with the results in Table 3.6 and Table 3.7 that the CLSA outperforms ACF or CCF based methods by 20+% in both rank-1 and mAP.

We further tested the person search effect of our detection model by comparing with the results based on ground-truth bounding boxes. It is found in Figure 3.8 (b) with perfect person detection, the CLSA gives only a gain of 0.9% (88.1-87.2) in mAP and 1.5% (90.0-88.5) in Rank-1. This indicates that the person detection component is not necessarily a major performance bottleneck in person search, thanks to modern object detection models. On the other hand, Table 3.8 also shows that addressing the multi-scale challenge is more critical for the overall model

Table 3.10: Effect of in-network feature pyramid construction on CUHK-SYSU.

Blocks Selection	5-4	5-4-3	5-4-3-2
Rank-1 (%)	87.3	**88.5**	85.3
mAP (%)	86.2	**87.2**	84.3

Table 3.11: Effect of temperature softness (Eq. (3.9)) on CUHK-SYSU.

Temperature T	1	3	5	7
Rank-1 (%)	88.3	**88.5**	88.3	88.1
mAP (%)	87.0	87.2	**87.3**	86.9

performance on person search, e.g. CLSA brings a performance boost of 6.0% (88.5-82.5%) in Rank-1 and 5.6% (87.2-81.6) in mAP over the baseline network ResNet-50.

Effect of Feature Pyramid We evaluated the performance effect of feature pyramid of CLSA on CUHK-SYSU. Recall that the in-network feature pyramid construction is based on the selection of ResNet blocks (see Sec. 3.2.1 and Figure 3.5). We tested three block selection schemes: 5-4, 5-4-3 (used in the final CLSA solution), and 5-4-3-2. Table 3.10 shows that a three-level pyramid is the optimal. It also suggests that performing semantic alignment directly with elementary features such as those extracted from the Res2X block may degrade the overall representation benefit in the pyramid, due to the hard-to-bridge semantic gap.

Effect of Temperature Softness We evaluated the impact of the temperature parameter setting in Eq. (3.9) in the range from 1 to 7. Table 3.11 shows that this parameter is not sensitive with the best value as 3.

Evaluating Person Re-ID and Object Classification We evaluated the effect of CLSA on person re-id (Market1501 [43], CUHK03 [40]) and object image classification (CIFAR100 [81]), in comparison to ResNet-50. Table 3.12 shows the positive performance gains of our CLSA method on both tasks. For example, the CLSA improves person re-id by 3.5%(88.9-85.4) in Rank-1 and 4.5% (73.1-68.6) in mAP on Market-1501. This gain is smaller than that on the same source video based PRW (see Table 3.7), due to the potential reason that person bounding boxes of Market-1501 have been manually processed with limited and artificial scale variations. Moreover, our method also benefits the CIFAR object classification with a 1.5% (76.2-74.7) top-1 rate gain. These observations suggest the consistent and problem-general advantages of our model in addition to person search in unconstrained scene images.

Table 3.12: Evaluating the CLSA on re-id and object classification benchmarks.

Dataset	Market-1501 [43]		CUHK03 [40]		Dataset	CIFAR100 [81]
Metric (%)	Rank-1	mAP	Rank-1	mAP	Metric (%)	Top-1 rate
ResNet-50	85.4	68.6	48.8	47.5	ResNet-110	74.7
CLSA	**88.9**	**73.1**	**52.3**	**50.9**	CLSA	**76.2**

3.3 Summary

The chapter investigates the weakness of cross-class knowledge transfer in person re-id. Specifically, the negative influence of background noise and multi-scale person bounding box to knowledge transfer are discussed. To solve the issue of background noise, we present an Identity DiscriminativE Attention reinforcement Learning (IDEAL) model for optimising re-id attention selection in auto-detected bounding boxes. This improves notably person re-id accuracy in a fully automated process required in practical deployments. The IDEAL model is formulated as a unified framework of discriminative identity learning by a deep multi-class discrimination network and attention reinforcement learning by a deep Q-network. This achieves jointly optimal identity sensitive attention selection and re-id matching performance by a reward function subject to identity label pairwise constraints. Extensive comparative evaluations on two auto-detected re-id benchmarks show clearly the advantages and superiority of this IDEAL model in coping with bounding box misalignment and background clutter removal when compared to the state-of-the-art saliency/attention based re-id models. Moreover, this IDEAL automatic attention selection mechanism comes near to be equal to human manual labelling of person bounding boxes on re-id accuracy, therefore showing great potential for scaling up automatic re-id to large data deployment. Besides, to alleviate the issue of multi-scale person bounding box to cross-class knowledge transfer, this chapter further proposes an end-to-end CLSA deep learning method by constructing an in-network feature pyramid structural representation and enhancing its representational power with a semantic alignment learning loss function. This is designed specially to make all feature pyramidal levels identity discriminative therefore leading to a more effective hierarchical representation for matching person images with large and unconstrained scale variations. Extensive comparative evaluations have been conducted on two large person search benchmarking datasets CUHK-SYSU and PRW. The results validate the performance superiority and advantages of the proposed CLSA model over a variety of state-of-the-art person search, person re-id and multi-scale learning methods. We also provide comprehensive in-depth CLSA component evaluation and analysis to give the insights on model performance gain and design considerations. In addition, we further validate the more general performance advantages of the CLSA method on the person re-identification and object categorisation tasks.

Chapter 4

Knowledge Transfer Across Domains in Person Re-identification

All knowledge is connected to all other knowledge. The fun is in making the connections.

—— **Arthur C. Aufderheide**

This chapter explores the cross-domain knowledge transfer in person re-identification. Specifically, two scenarios are discussed. First of all, unsupervised domain adaptation for person re-id is investigated. A novel Hierarchical Unsupervised Domain Adaptation (HUDA) is proposed to promote the effectiveness of cross-domain knowledge transfer. In particular, HUDA is designed to strengthen the local instance alignment, which is ignored by existing work. Second, this chapter provides a novel universal re-id approach to enable the trained model to deploy any target domains, addressing the weakness of the cross-domain knowledge transfer can only be performed well in the single target domain deployment.

4.1 Unsupervised Person Re-id for Domain Adaptation

4.1.1 Unsupervised Hierarchical Adaptation

Problem statement. For unsupervised cross-domain person re-id, we have a *supervised* (labelled) source dataset (domain) $D^s = \{\boldsymbol{I}_i^s, y_i^s\}_{i=1}^{K^s}$, consisting of K^s person bounding box images $\boldsymbol{I}_i^s$ each with the corresponding *identity* label $y_i^s \in \mathcal{Y} = \{1, \cdots, K_{\text{id}}^s\}$, i.e. a total of K_{id}^s different persons in the source domain. Meanwhile, we assume a set $D^t = \{\boldsymbol{I}_i^t\}_{i=1}^{K^t}$ of K^t *unsupervised* (unlabelled) training data randomly sampled from the target domain with unknown and non-overlapping identity labels. Using D^t is for model domain adaptation. The ***goal*** is to learn a feature representation optimal for the unlabelled target domain ID class discrimination by transferring the identity discriminative information learned from a labelled source domain.

Approach overview. We present a *Hierarchical Unsupervised Domain Adaptation* (HUDA) model. It can jointly perform *global feature distribution alignment* and *local instance alignment* between the source and target domains by end-to-end deep learning. This is uniquely characterised by *more fine-grained* knowledge transfer during unsupervised domain adaptation. This is crucial for person re-id since a key objective is to capture subtle discrimination of different persons with high appearance similarity. Aligning only global distributions across domains is insufficient due to being highly under-constrained. With a joint modelling, fine-grained instance alignment enriches global distribution alignment. An overview of HUDA is depicted in Figure 4.1.

Person Re-Identification Model

To build a re-id model $\boldsymbol{\theta}^{\text{tar}}$ (Figure 4.1(c_1, c_2)), we use ResNet-50 [5] as backbone. Given labelled *source* training data D^s, we train the model by a discriminative loss function $\mathcal{L}_{\text{re-id}} = \mathcal{L}_{\text{ce}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}}$ where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{tri}}$ denote the softmax Cross Entropy loss (Eq (3.7)) and the triplet loss, respectively. We empirically set the weight parameter $\lambda_{\text{tri}} = 0.3$.

Discussion. A trained re-id model by the above formulation is suitable *only* for the source domain deployment, therefore having limited generalisation. To adapt the model to an independent target domain, we perform unsupervised domain adaptation by a HUDA model. In HUDA, *unlabelled* target domain data are used as a bridge for transferring source domain knowledge. Our model consists of two parts: (1) global distribution alignment, and (2) local instance alignment.

Global Distribution Alignment

The Global Distribution Alignment (GDA) component of HUDA aims to adapt holistic statistical information between the source and target domains (Figure 4.1(d)). Due to the disjoint nature of source and target identity classes (i.e. an open-set recognition setting), GDA seems improper and has been shown to be ineffective for generic open-set object classification [222, 223]. Nonetheless, person re-id is rather different from generic object recognition, since it is a fine-grained matching problem.

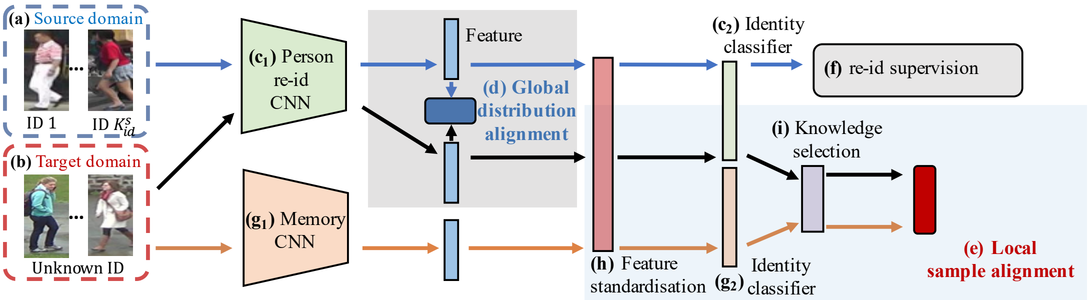


Figure 4.1: Overview of *Hierarchical Unsupervised Domain Adaptation* (HUDA). Given **(a)** *supervised* source domain and **(b)** *unlabelled* target training person imagery data, we aim to learn **($\mathbf{c_1}$, $\mathbf{c_2}$)** a re-id model generalisable to the target domain. To this end, the proposed HUDA model jointly conducts **(d)** Global Distribution Alignment (GDA) and **(e)** Local Instance Alignment (LIA) in an end-to-end network learning architecture subject to **(f)** source re-id supervision. Cross-domain adaptation by the GDA alone is highly under-constrained. We address this by introducing the LIA for more fine-grained unsupervised domain adaptation with the stronger constraint. In re-id, there is often no identity class overlap between the source and target domains. Motivated by our primitive attribute viewpoint, we leverage cross-class association to discover and exploit *reliably transferable* knowledge for domain adaptation. This is achieved by the proposed LIA through incrementally building **($\mathbf{g_1}$, $\mathbf{g_2}$)** a knowledge memory network to cumulatively memorise the past learned knowledge throughout training and simultaneously offer target domain instance-specific local knowledge for high quality adaptation from the labelled source domain to the unlabelled target domain. To further improve the knowledge quality, we introduce **(h)** a feature standardisation layer to accelerate the model training and **(i)** a knowledge selection mechanism for more reliable domain adaptation.

A counter-intuitive phenomenon in re-id. Essentially, person re-id aims to derive a feature representation for pairwise similarity based matching and ranking. The training and testing person identity classes are totally *disjoint*. Such *cross-class* (i.e. open-set recognition) nature between training and testing is *universal* and *intrinsic* to the problem. Consider that the learning target is for optimal *pairwise matching*, early deep re-id models reasonably use *pairwise loss functions* (including the triplet ranking loss involving positive and negative pairs) for model training [40, 224, 31]. Subsequent works empirically find that the softmax Cross-Entropy (CE) loss, which is commonly used for training *closed-set* multi-class classification models, is similarly effective, even without the complexity of pairing samples [29]. This selection (presumably occasional) is actually *not* as intuitive as the pairwise counterparts, because the CE loss is conventionally considered effective *only* for *closed-set* recognition [225], so it would have been "*ineffective*" for cross-class learning as re-id. That being said, this traditional wisdom is *against* the wide practices. Interestingly, this counter-intuitive phenomenon lacks proper interpretation in the literature.

The essence to cross-class recognition in re-id. We provide an explanation to the above phenomenon as follows. By learning re-id feature representation for pairwise similarity matching, we consider the *fundamental key* is to derive *a set of primitive patterns (attributes)* which are formally composited of individual feature dimensions or some dimension combinations. They are useful to distinguish different person appearance and largely *independent* of any person identity classes including training classes. That is, these primitive attributes can describe arbitrary person appearance due to their massive combination space, which is the *essence* for them to possess cross-class recognition capability. Therefore, the essential learning objective is to obtain such a set of class independent primitive attributes, rather than a pairwise similarity matching function (previous understanding). Consequently, it is not necessarily to limit the learning objective to pairwise loss functions; The CE loss function can be similarly effective since the learning of classifiers also results in a set of primitive attributes optimal for multi-class discrimination. These loss functions are *functionally* similar in this primitive attribute viewpoint. This naturally interprets the *mysterious* efficacy of the CE loss for re-id.

Cross-domain in re-id. Unlike the generic object class classification with distinct appearance difference [223, 222], person re-id handles uniquely fine-grained identity discrimination with similar holistic person appearance. This suggests that a large proportion of primitive attributes can be shared across domains, i.e. overlapped in the distribution. Specifically, the feature representations contain more primitive attributes shared over domains. Together with cross-class interpretation, GDA navigates cross-domain person re-id learning.

GDA formulation. Due to highly complex distributions of visually ambiguous and diverse re-id image data, it is difficult to select a suitable parametric model for such a distribution. We adopt a non-parametric representation to characterising re-id visual data statistics. In particular, we exploit the Maximum Mean Discrepancy (MMD) [226] to measure the feature dissimilarity between the source and target domains for distribution alignment:

$$\mathcal{L}_{\text{mmd}^2} = ||\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\boldsymbol{f}_{s,i}) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(\boldsymbol{f}_{t,j})||_{\mathcal{H}}^2 \tag{4.1}$$

where $\boldsymbol{f}_s \in \mathbb{R}^{n_s \times d}$ and $\boldsymbol{f}_t \in \mathbb{R}^{n_t \times d}$ specify the feature vectors of n_s source and n_t target images in each mini-batch, and d is the feature dimension. We further enforce non-linearity by using a mapping function $\phi(\cdot)$ to project the feature samples into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ [227]. By the kernel trick, we design the GDA loss by reformulating Eq (4.1) as:

$$\begin{aligned} \mathcal{L}_{\text{gda}} = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{i'=1}^{n_s} k(\boldsymbol{f}_{s,i},\boldsymbol{f}_{s,i'}) + \\ \frac{1}{n_t^2}\sum_{j=1}^{n_t}\sum_{j'=1}^{n_t} k(\boldsymbol{f}_{t,j},\boldsymbol{f}_{t,j'}) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\boldsymbol{f}_{s,i},\boldsymbol{f}_{t,j}) \end{aligned} \tag{4.2}$$

We adopt the common Gaussian kernel function:

$$k(\boldsymbol{f}_{s,i},\boldsymbol{f}_{t,j}) = \exp\Big(-\frac{||\boldsymbol{f}_{s,i}-\boldsymbol{f}_{t,j}||_2^2}{2\sigma^2}\Big) \tag{4.3}$$

where σ is the kernel bandwidth. To reduce the selection bias and enable to automatically identify an optimal kernel, we deploy a predefined set of kernels with $\sigma \in \{1,5,10\}$.

Local Instance Alignment

To enrich GDA based cross-domain adaptation by cross-class discriminative learning necessary for person re-id, we further introduce Local Instance Alignment (LIA) to explore instance level fine-grained discriminative learning (Figure 4.1(e)). Specifically, we want to progressively discover and adapt *reliably transferable* source information specific to individual target samples during training. The key idea is learning to associate target samples with visually similar source data for guiding cross-domain knowledge transfer. The intuition is that, re-id of target instances can benefit ("borrow" information) from a model discriminatively trained by labelled source instances if the target and source instances are visually aligned (similar).

The association in LIA is often across identity classes between domains. Inspired by our primitive attribute viewpoint, we classify the target person images into the source identity classes. Specifically, given an unlabelled target person image sample $\boldsymbol{I}^t$, we predict a class probability vector for it in the source domain class-label space:

$$\boldsymbol{p}(\boldsymbol{I}^t) = \{p(1|\boldsymbol{I}^t), p(2|\boldsymbol{I}^t), \cdots, p(K_{\mathrm{id}}^{\mathrm{s}}|\boldsymbol{I}^t)\} \tag{4.4}$$

This classification indicates how visually similar a target person image is measured against all the source classes. It encodes the *cross-domain transferable knowledge* we aim to extract for unsupervised domain adaptation.

Source Knowledge Discovery

In a unified design, the source and target domain model learning shares a single network trained *simultaneously*. A faster training on the source data is essential for ensuring the knowledge quality. Consider deep learning using mini-batches of training samples as a stochastic learning process, the feature distribution changes per batch. This may complicate and slow down the unsupervised domain adaptation process, because the model needs to repeatedly and continuously adapt to new distributions throughout the training process.

Feature standardisation. To address the above problem, we enforce that the model always outputs the feature representations in a fixed distribution. Specifically, we standardise the re-id feature representations (the average pooling of the last conv layer of ResNet-50). This performs a per-dimension normalisation on the per-batch feature vectors from both domains (Figure 4.1(h)), as follows:

$$\hat{\boldsymbol{f}} = \frac{\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}]}{\sqrt{\mathbb{V}[\boldsymbol{f}] + \varepsilon}} \tag{4.5}$$

where $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote the per-dimension expectation and variance of feature values per batch. The small constant $\varepsilon > 0$ is for ensuring numerical stability. Given this, we use the standardised features $\hat{\boldsymbol{f}}$ for re-id deployment in test.

Remarks. Feature standardisation has been used elsewhere, e.g. Sparsifying Features [228], and Batch Normalisation (BN) [229]. In this study, we investigate its potential for unsupervised domain adaptation in person re-id. The key differences are: Compared to BN that introduces two extra free parameters for scaling and shift in order to preserve the identity transform respectively, our method does not have such requirements. BN is used to normalise the layer inputs, whereas our model is applied to the model output. In contrast to [228], our method does not improve the feature sparsity nor constrain the internal layer outputs.

Knowledge memory network. To project a target instance into the source identity class space, a straightforward way is to apply the current up-to-date deep model. However, this is not ideal. The reason is as follows. In stochastic deep learning, the in-training model updates at each iteration. This may cause the model performance to temporally deteriorate on samples of the past mini-batches, due to the nature of *catastrophic forgetting* [230]. As target domain samples are randomly sampled, it is possible that the up-to-date model has degraded in recent updates when assessing some target samples of the current batch.

To further improve the knowledge quality, we propose to incrementally memorise the source information learned per mini-batch during training. In particular, we establish a *knowledge memory network* (Figure 4.1(g_1,g_2)) $\boldsymbol{\theta}^{\text{mem}}$ in identical architecture as the target model, and we exploit it to obtain the knowledge in the form of class posterior probability. Formally, this knowledge memory network $\boldsymbol{\theta}^{\text{mem}}$ is updated along with the target model $\boldsymbol{\theta}^{\text{tar}}$ at each iteration τ by exponential moving average as:

$$\boldsymbol{\theta}_{\tau}^{\text{mem}} = \alpha\boldsymbol{\theta}_{\tau-1}^{\text{mem}} + (1-\alpha)\boldsymbol{\theta}_{\tau}^{\text{tar}} \tag{4.6}$$

where α is the smoothing coefficient hyper-parameter. We set $\alpha = 0.99$ empirically. In doing so, the discriminative information derived from each mini-batch is absorbed and memorised into $\boldsymbol{\theta}^{\text{mem}}$, so that the memory model serves as a stronger knowledge extractor as compared to the up-to-date target model. That is, in mini-batch training we exploit the $\boldsymbol{\theta}_{\tau}^{\text{mem}}$ as the replacement of $\boldsymbol{\theta}_{\tau}^{\text{tar}}$ to obtain the posterior probability vector (Eq (4.4)) for each unlabelled target sample in the source domain class space.

Remarks. The proposed memory network is inspired by the neuron memory mechanism [231]. This is due to that the memorising capacity of deep networks is often incomplete and limited in representing knowledge experienced in the past learning iterations. However, unlike [231], our method uses a network for memory organisation without the need for extra components to customise the network structure and designing particular knowledge representations for access operations. Algorithmically, building our memory network is similar to the notion of mean-teacher in semi-supervised learning [76], but the two address different goals. Our method seeks a reliable cross-class knowledge extraction in training. In contrast, mean-teacher aims to improve label prediction on unlabelled data from the same domain in a closed-set classification setting.

Source Knowledge Transfer

The aim of source knowledge transfer is to enhance the generalisation of the target model $\boldsymbol{\theta}^{\text{tar}}$ in the target domain. To this end, we consider the richer memorised knowledge in the memory network that is relevant to target domain samples. However, the underlying transferable knowledge between source and target domains is *unknown* a priori. It is sub-optimal to blindly transfer all memory knowledge with all target samples. To address this, we design a knowledge selection mechanism (Figure 4.1(i)) for more reliable adaptation on individual samples.

Knowledge selection. In unsupervised cross-domain re-id, not all target person images can be associated with some source identity classes with high confidence. This is due to the cross-class nature between independent domains with entirely different person classes. Given that source knowledge is expressed in a probability form, one intuitive way to measure the knowledge transferability and reliability is to use the maximum likelihood:

$$\mathcal{ML}(\boldsymbol{I}^t) = \max(\{p(1|\boldsymbol{I}^t), p(2|\boldsymbol{I}^t), \cdots, p(K^{\text{s}}_{\text{id}}|\boldsymbol{I}^t)\}) \tag{4.7}$$

With this, we can then deploy a thresholding strategy for knowledge selection by choosing those target samples satisfying that the corresponding $\mathcal{ML}(\boldsymbol{I}^t)$ exceeds a pre-defined threshold u. We denote the selected target samples as $\tilde{\boldsymbol{I}}^t$. In cross-class context, it is often that most $\mathcal{ML}(\boldsymbol{I}^t)$ values are not high. Hence, a mild threshold value is preferred to ensure sufficient source-target associations. Too small threshold values, on the other hand, may lead to adapting non-transferable knowledge with negative effects. We empirically find that setting $u = 0.3$ is satisfactory.

Knowledge transfer. Once we have the selected knowledge, the next is to transfer it into the target model, i.e. knowledge domain adaptation. To accomplish this, we align the knowledge memory model and the target model in their predictions of selected target samples $\tilde{\boldsymbol{I}}^t$ by exploiting the Kullback-Leibler (KL) divergence (Eq (2.4)) written as:

$$\mathcal{L}_{\text{lia}} = \sum_{j=1}^{K^{\text{s}}_{\text{id}}} p_(j|\tilde{\boldsymbol{I}}^t, \boldsymbol{\theta}^{\text{mem}}) \log \frac{p_(j|\tilde{\boldsymbol{I}}^t, \boldsymbol{\theta}^{\text{mem}})}{p_(j|\tilde{\boldsymbol{I}}^t, \boldsymbol{\theta}^{\text{tar}})} \tag{4.8}$$

Overall Model Loss Formulation

Given the re-id and HUDA loss functions, we obtain the final objective function for model training as:

$$\mathcal{L} = \mathcal{L}_{\text{re-id}} + \lambda_{\text{gda}}\mathcal{L}_{\text{gda}} + \lambda_{\text{lia}}\mathcal{L}_{\text{lia}} \tag{4.9}$$

where λ_{gda} and λ_{lia} are the relative importance parameters. We set $\lambda_{\text{gda}} = 1$ and $\lambda_{\text{lia}} = 1$ in our experiments. The whole model can be trained end-to-end subject to the loss function of Eq. (4.9)

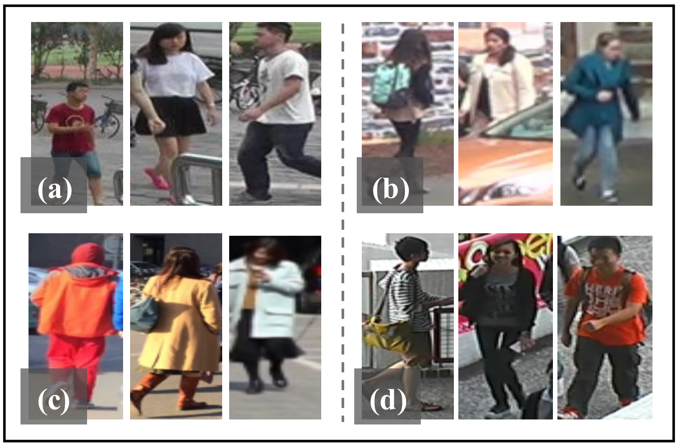


Figure 4.2: Example person images from (a) Market-1501, (b) DukeMTMC, (c) CUHK03, (d) MSMT-17.

by the stochastic gradient descent algorithm.

4.1.2 Experiments

Datasets. For evaluation, We used four person re-id benchmarks with distinct camera viewing conditions. (Figure 4.2). The ***Market-1501*** [43] contains 32,668 images of 1,501 identities (ID) captured by 6 cameras. We used the standard 751/750 train/test ID split. The ***DukeMTMC*** [48, 59] consists of 36,411 labelled images of 1,404 IDs from 8 camera views. We adopted the same 702/702 ID split as [48]. The ***CUHK03*** [40] provides 14,096 images of 1,467 IDs from 6 camera views. We used the detected images as the source as [50]. The ***MSMT-17*** [41] is a largest person re-ID benchmark thus far. contains 126,411 person images from 4,101 IDs captured from 15 camera views. We adopted the standard 1041/3060 train/test ID split.

Performance metrics. We adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the model performance measurements.

Model parameter setting. In this context, no target domain supervision is available for hyper-parameter cross-validation. We hence used a *single* set of empirical parameter setting for HUDA (including λ_{tri} for $\mathcal{L}_{\text{re-id}}$, α in Eq (4.6), u for Eq (4.7), λ_{gda} and λ_{lia} in Eq (4.9)) in all the experiments.

Implementation details. The backbone ResNet-50 was pre-trained on ImageNet. To train a re-id model, we deployed SGD with the momentum set to 0.9, the weight decay to 0.0005, and the mini-batch size of 64 (32 source plus 32 target samples), the epoch number to 60. All input images were resized to 256×128 and subtracted by ImageNet mean. We applied data augmentation for the target and memory networks independently in training, including random cropping, random flipping, and colour jitter. In test time, we used the Euclidean distance as the re-id matching metric.

Comparisons to the State-of-the-Art Methods

For a fine-grained evaluation, we compared five types of existing methods: **(a)** two hand-crafted feature models, (LOMO [42], BoW [43]); **(b)** three image adaptation models (PTGAN [41], SPGAN+LMP [49], ATNet [51]), **(c)** six feature adaptation models (UMDL [52], CAMEL [155], PUL [232], TJ-AIDL [53], MMFA [54], MAR [55]), **(d)** one unsupervised deep learning method (TAUDL [56]), **(e)** three hybrid methods (HHL [50], ECN [57], PAUL [58]). We evaluated three transfer scales in terms of the source data size: (1) large: MSMT17⇒Market, (2) medium: Market1501⇔DukeMTMC, (3) small: CUHK03⇒Market.

Table 4.1: Results on Market-1501⇔DukeMTMC.

Source→Target	Duke→Market				Market→ Duke			
Metric (%)	R1	R5	R10	mAP	R1	R5	R10	mAP
LOMO [42]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
BOW [43]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
PTGAN [41]	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN+LMP [49]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
ATNet[51]	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
TAUDL [56]	63.7	-	-	41.2	61.7	-	-	43.5
UMDL [52]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
CAMEL [155]	54.5	-	-	26.3	-	-	-	-
PUL [232]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
TJ-AIDL [53]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
MMFA [54]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
HUDA(Ours)	68.5	82.9	87.1	37.6	52.3	65.4	68.7	30.2
HHL[50]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
ECN [57]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
HUDA+TAUDL	**78.8**	**90.2**	**93.4**	**57.6**	**70.4**	**82.5**	**86.2**	**51.2**

Evaluation on DukeMTMC ⇔ Market-1501. Table 4.1 shows the comparisons between HUDA and 13 state-of-the-art methods. We have the following observations. **(1)** Hand-crafted feature methods [42, 43] produce the poorest performance, due to weak representations. **(2)** Image adaptation methods [41, 49, 50] yield fairly strong re-id rates. **(3)** Interestingly, the unsupervised tracklet re-id method [56] achieves competitive performance without using any labelled source data. **(4)** For the feature adaptation models, HUDA outperforms all the competitors [52, 155, 232, 53, 54]. This suggests strongly the modelling superiority of our method over the state-of-the-art models. **(5)** For like-to-like comparisons with hybrid methods, we combined HUDA with TAUDL in a two-stage process – using HUDA for model pre-training then TAUDL for unsupervised target data learning. This hybrid model, called *HUDA+TAUDL*, achieves the best results as compared to HHL and ECN.

Table 4.2: Results on MSMT17/CUHK03⇒Market-1501.

S→T	MSMT→Market				S→T	CUHK→Market			
Metric(%)	R1	R5	R10	mAP	Metric(%)	R1	R5	R10	mAP
MAR	67.7	81.9	-	40.0	HHL[50]	42.7	57.5	64.2	23.1
PAUL [58]	68.5	82.4	87.4	40.1	SPGAN [49]	42.3	-	-	19.0
HUDA	**72.3**	**85.2**	**89.2**	**42.4**	**HUDA**	**49.7**	**62.8**	**67.7**	**27.9**

Evaluation on MSMT17/CUHK03 ⇒ Market-1501. We further tested the domain adaptation with large and small scale transfer. Table 4.2 compares the performance of HUDA to 4 state-of-the-art alternative methods with reported re-id results. Overall, we have similar observation as above. For MSMT17⇒Market-1501, as a feature adaptation method, HUDA even surpasses the hybrid competitor PAUL. In the case of small scale transfer on CUHK03⇒Market-1501, HUDA consistently outperforms all strong competitors. This test validates the superiority of HUDA in varying cross-domain adaptation scenarios.

Further Analysis and Discussions

We conducted a series of component analysis for HUDA using DukeMTMC⇔Market-1501.

HUDA design. We tested the significance of HUDA and its components (GDA and LIA). Table 4.3 shows that: (1) Without HUDA, the model suffers clearly the domain gap, e.g. large performance drop. (2) GDA *Only* gives significant performance boost. This validates our *primitive*

Table 4.3: HUDA design analysis. GDA: Global Distribution Alignment. LIA: Local Instance Alignment.

Source→Target	Duke→ Market				Market→Duke			
Metric(%)	R1	R5	R10	mAP	R1	R5	R10	mAP
w/o HUDA	55.2	74.3	81.3	27.1	41.8	57.6	63.2	22.3
GDA Only	61.8	77.9	83.6	32.4	46.8	62.6	68.8	26.5
LIA Only	61.9	78.3	83.8	32.9	44.3	59.4	65.5	24.1
Full HUDA	**68.5**	**82.9**	**87.1**	**37.6**	**52.3**	**65.4**	**68.7**	**30.2**

attribute interpretation . (3) LIA *Only* also yields similar re-id rate gain. This verifies the idea of our local alignment and the proposed design. (3) When GDA and LIA are jointly exploited (i.e. full HUDA), model performance is further increased. This validates good complementary of GDA and LIA, as well as our motivation of integrating them into a single formulation.

Cross-class association between domains. Recall that we classify the unlabelled target person images into the source identity class space in a cross-class manner. This aims to associate target persons with visually similar source people in the LIA process (see Figure 4.4). We examined the effectiveness of this association. Specifically, we measured the proportion of target person images highly associated to any source identity classes with the maximum likelihood above the threshold u. We tracked this measurement *with* and *without* the LIA. We observed from Figure 4.3 that, the proposed association scheme significantly improves the cross-domain alignment at the fine-grained instance level. LIA makes the most target persons associated to the relevant source identities with similar appearance. This indicates that GDA is *under-constrained.* Not every target sample can be associated with a visually similar source identity by HUDA. This is reasonable due to the independent nature between source and target domains. The rising association rate of HUDA *without* LIA in the beginning of training is due to inaccurate predictions by the *immature* in-training model.

Feature standardisation. We evaluated the effect of feature standardisation (FS) on unsupervised domain adaptation *with* and *without* HUDA. Table 4.4 shows that FS is significant for effective cross-domain knowledge transfer in HUDA context, validating our design consideration. This is because, the cross-domain association becomes reliable and effective for unsupervised domain adaptation, only when the model learns sufficiently discriminative information from the

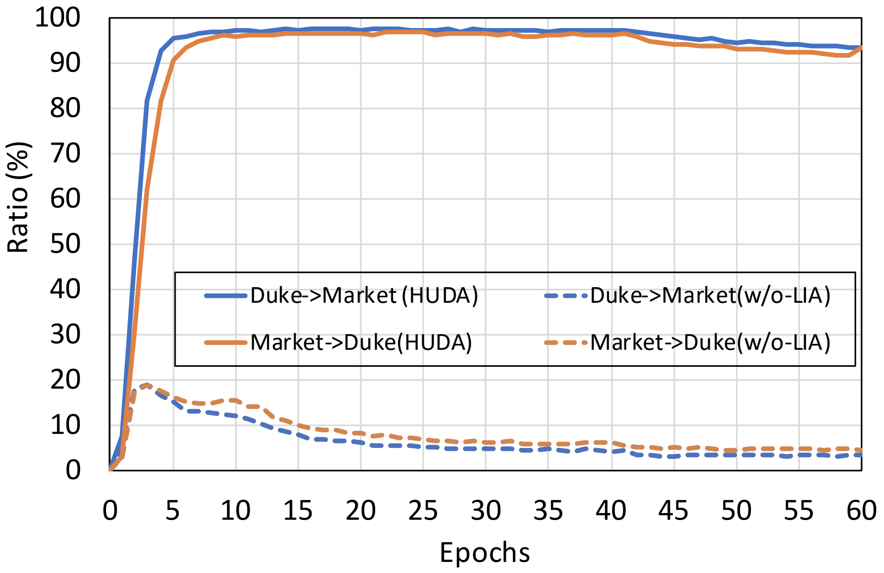


Figure 4.3: The proportion of target training samples that is highly associated with source classes during model training.

source labels. Besides, FS slightly helps the baseline without HUDA, suggesting a generic usefulness. We further tested the impact of FS on the model performance convergence on the source domain data. We chose the memory network that is used for knowledge extraction. Figure 4.5 shows that FS is clearly beneficial for accelerating the model learning speed on the source labelled data.

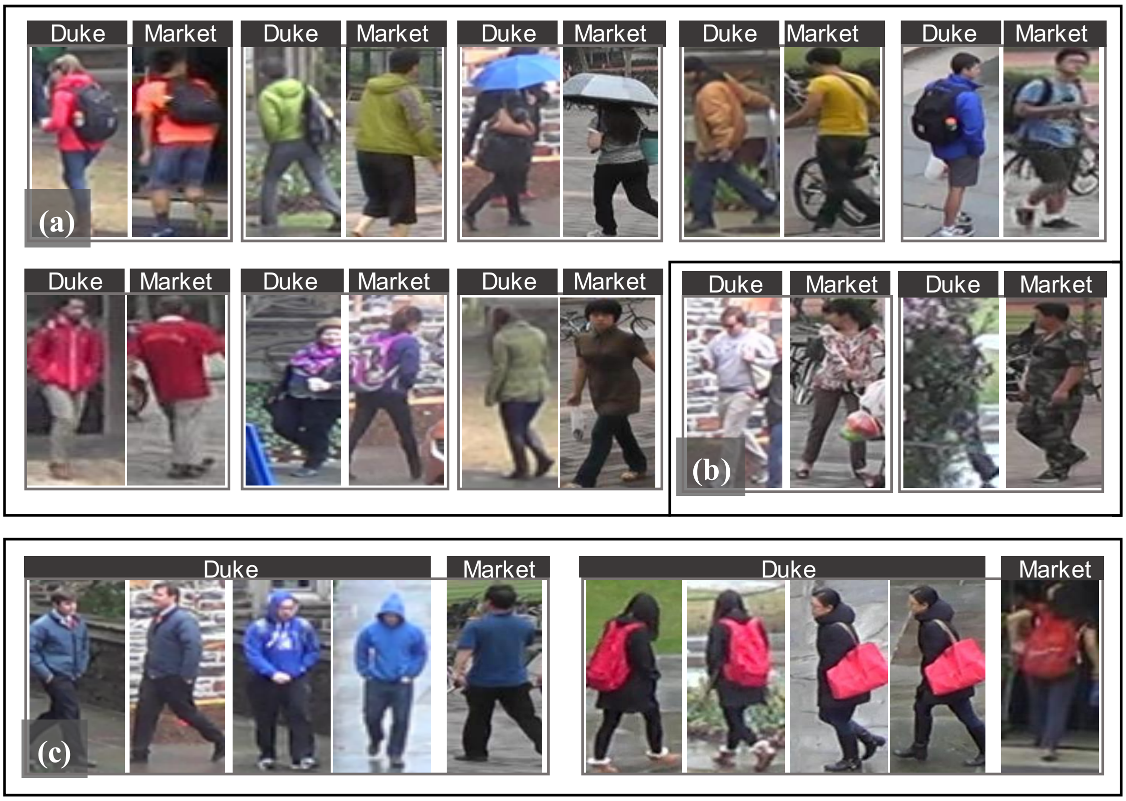


Figure 4.4: Association of *target* DukeMTMC persons to *source* Market-1501 identity classes. **(a)** The pairs of source and target persons extracted automatically by cross-domain cross-class association. The associated persons show strong visual similarities. **(b)** The target person images associated to a source person have either the *same* identity (when in the same domain) or *similar* visual appearance (when cross-domain). **(c)** Cross-domain associations can be distracted by background clutters.

Table 4.4: Examination of feature standardisation (FS).

Source→Target		Duke→ Market				Market→Duke			
Model	FS	R1	R5	R10	mAP	R1	R5	R10	mAP
w/o HUDA	✗	55.2	74.3	**81.3**	27.1	41.8	57.6	63.2	22.3
w/o HUDA	✓	**56.9**	**74.2**	80.1	**28.4**	**42.1**	**57.9**	**63.3**	**22.5**
HUDA	✗	61.5	77.2	82.9	32.3	44.5	57.6	64.0	24.6
HUDA	✓	**68.5**	**82.9**	**87.1**	**37.6**	**52.3**	**65.4**	**68.7**	**30.2**

Table 4.5: Examination of knowledge selection (KS).

Source→Target	Duke→ Market				Market→Duke			
KS	R1	R5	R10	mAP	R1	R5	R10	mAP
✗	65.5	79.1	84.7	34.7	48.3	63.5	67.9	27.5
✓	**68.5**	**82.9**	**87.1**	**37.6**	**52.3**	**65.4**	**68.7**	**30.2**

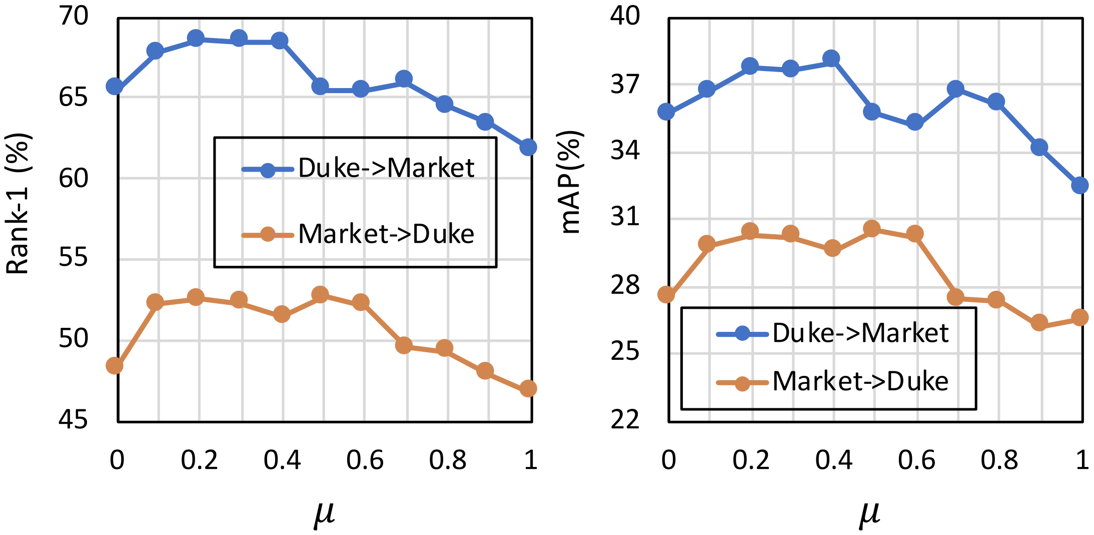


Figure 4.6: Effect of controlling the knowledge reliability in cross-domain transfer in (left) Rank-1 and (right) mAP rates.

Knowledge selection. We tested the performance benefit from knowledge selection (KS). The KS is controlled by setting a threshold u on the maximum likelihood in the source class space (Eq (4.7)). We compared the re-id accuracy rates on the target domain *with* and *without* the

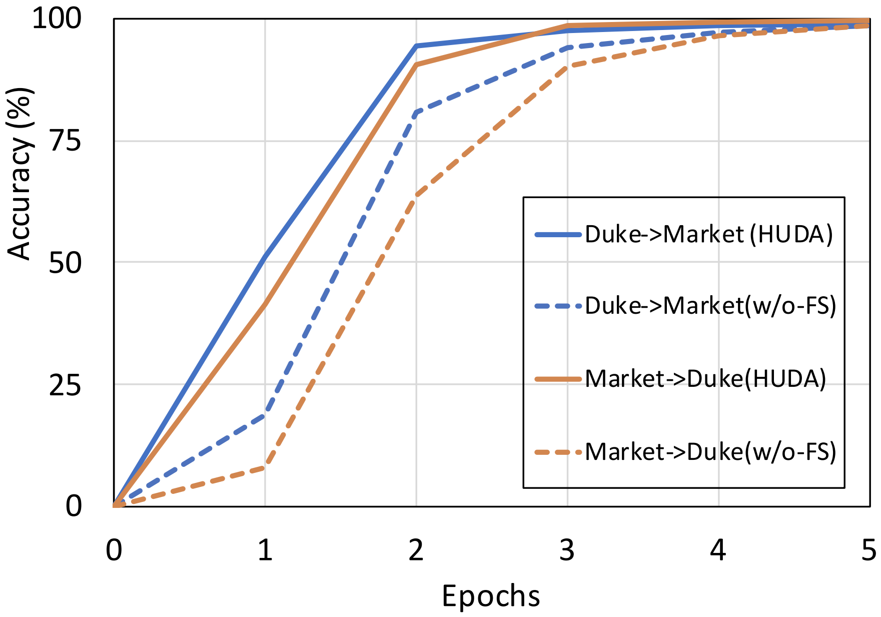


Figure 4.5: Effect of the feature standardisation (FS) to the model convergence on the source domain data.

thresholding based (u) selection. Table 4.5 and Figure 4.6 support the significance of knowledge selection for more reliable unsupervised domain adaptation. The optimal selections lie in the range of $[0.1, 0.4]$, validating our consideration that a mild threshold value u would be used. Note that not the *entire* ($u = 0$) source knowledge are equally relevant and reliably transferable to the target domain. Adapting unsuitable source information can hurt the model generalisation. Besides, the performance is clearly inferior when *no* local knowledge adaptation is considered ($u = 1$), validating our modelling motivation.

Table 4.6: Domain adaptation (DA) effects on the source domain.

Dataset	Market				Duke			
Metric(%)	R1	R5	R10	mAP	R1	R5	R10	mAP
Before DA	86.6	94.7	97.0	67.5	**77.4**	**88.5**	**91.7**	**59.5**
SPGAN[49]	59.9	78.7	84.5	34.3	53.9	70.9	76.5	32.4
HUDA	**87.0**	**95.0**	**97.1**	**67.8**	77.1	87.9	91.4	59.3

Source domain performance. Unlike the image adaptation methods [49], HUDA avoids the need for re-id model fine-tuning for target domain. This helps maintain the model performance on the source domain. Table 4.6 shows that HUDA can preserve well model performance on the source data *after* domain adaptation. In contrast, SPGAN suffers significantly due to losing much of original discrimination ability in fine-tuning.

4.2 Universal Person Re-id

4.2.1 Domain-generic Image Transformation

To train a universal person re-id model, we assume a labelled source training dataset (i.e. a source domain) $\mathcal{I} = \{\boldsymbol{I}_i, y_i\}_{i=1}^{N}$, consisting of N person bounding box images $\boldsymbol{I}_i$ each annotated with a person identity class label $y_i \in \{1, \cdots, N_{\text{id}}\}$, It contains a total of N_{id} different person identities. We propose to transform this source training set D so as to cover the camera viewing conditions of arbitrary domains. Formally, we define a set of transformations $\{\mathcal{M}_{\boldsymbol{t}}\}, \boldsymbol{t} \in \mathcal{T}$ where $\boldsymbol{t}$ defines the transformation parameter vector and $\mathcal{T}$ represents the transformation space. Each transformation $\mathcal{M}_{\boldsymbol{t}}$ is composited of several primitive transformations. Considering the variations of person appearance at typical surveillance scenes are largely due to illumination (lighting), we establish a space of linear transformations with regard to pixel colour and contrast. Note, the approach is flexible to adopt other transformations if needed.

We consider the colour transformations in the HSV representation space [233]. Each colour has three fundamental attributes (Figure 4.7): (a) *Hue*: Colour such as red, orange, yellow, and so forth. It depends on the wavelength of light reflected and/or produced. (b) *Saturation* (*Chroma*): The brilliance of a hue, i.e. how pure (intense) a hue is. More saturated a hue is, brighter it appears. (c) *Lightness* (*Value*): The lightness or darkness of a hue. Adding white (black) makes the colour lighter (darker). Note, the effect of lightness is relative to other values in a composition.

Specifically, for hue transformation we convert the image into HSV and add the corresponding parameter to the original value on the hue dimension. Afterwards, the image is converted back. For the other factors (including contrast), we perform the transformation by linear interpolation and extrapolation [235]. For restricting the transformations to perceptually sensible scope, we define the variation range as: hue in [-18, 18] (cyclical), saturation in [0.6, 1.4], lightness in [0.6, 1.4], and contrast in [0.6, 1.4]. For saturation/lightness/contrast, the value of “1” means *no*

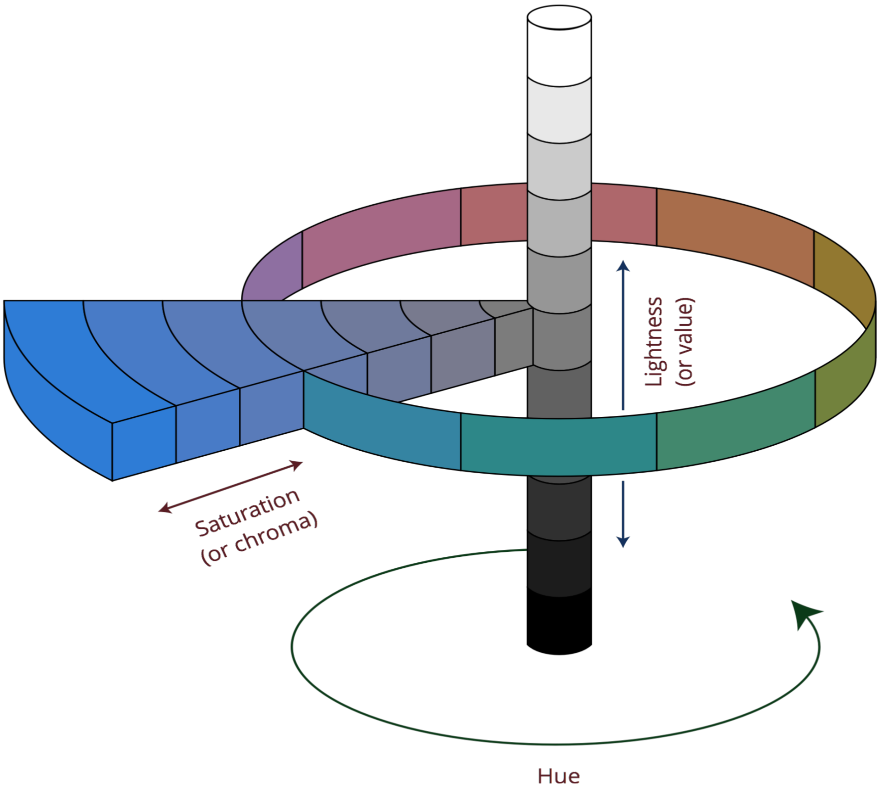


Figure 4.7: Illustration of Munsell colour system in three dimensions: Hue, Saturation, Lightness. This graph is adopted from [234]. Best viewed in colour.

transformation, and for hue “0” means *no* transformation.

To form a single colour-contrast transformation, we sample a parameter value for each factor and concatenate them into a vector $\boldsymbol{t}$. We consider an online image transformation strategy for stochastic deep learning. This avoids the need for saving and managing a large quantity of transformed images. In a training iteration, given a source image $\boldsymbol{I}_i$, we randomly sample a parameter vector $\boldsymbol{t}_i$ and apply the corresponding transformation $\mathcal{M}_{\boldsymbol{t}_i}$ to it. As such, we obtain a transformed variant:

$$D_i = \boldsymbol{I}_i \mathcal{M}_{\boldsymbol{t}_i}. \tag{4.10}$$

By repeating the transformation on each and every person image of a training mini-batch, we form domain generic universal training samples $\{D_i\}$ for model training on-the-fly. We show examples of transformed person images in Figure 4.8. Perceptually, such transformations leave the original identity class information of person images intact, facilitating the re-id discriminative model optimisation.

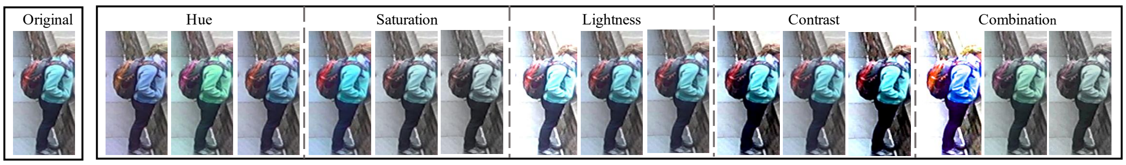


Figure 4.8: Example transformations in Hue, Saturation (Chroma), Lightness (Value), Contrast, and their random combinations.

Person Re-Identification Model

For person re-id model, we use ResNet-50 [5] as the backbone network. To enable fine-grained part-level discriminative learning, we adopt the PCB design [236]. Instead of the whole image, PCB uses average pooling on local regions and applies a separate re-id loss supervision on each individual region independently and concurrently. In addition, we add a parallel global branch for discriminative learning of the whole images. We apply label smoothing for mitigating model overfitting. For model training, we adopt the softmax Cross Entropy loss as the objective function:

$$\mathcal{L}_{\text{ce}} = -\sum_{k=1}^{N_{\text{id}}} \sum_{j=1}^{m} \delta_{k,y} \Big(\log \big(p_j(k|\boldsymbol{I}_i) \big) + \log \big(p(k|\boldsymbol{I}_i) \big) \Big) \tag{4.11}$$

where $\delta_{k,y}$ is the Dirac delta returning 1 if k is the ground-truth class label y, otherwise 0. p_j and p denotes the class posterior probability of the j-th local region and the whole image, estimated by the current network. m indicates the total number of local regions. We set $m = 6$ in our experiments the same as [236].

In test, we concatenate all the local regional and global features as the final re-id representation. We adopt the Euclidean distance metric for re-id matching and ranking.

Remarks

Compared to the previous data augmentation approaches [16, 3], our method differently focuses on training a universal model for any domain generalisation, other than enriching domain-specific training data variety and learning a better model for that domain alone. In particular, we uniquely consider training data transformations that simulate the person appearance distributions and characteristics of arbitrary unseen domains. Such data augmentation is not necessarily beneficial for the labelled training data domain. As the conventional wisdom suggests that ***colour is a key unique evidence of identity***, the data augmentation strategy used in existing re-id methods [34, 236, 237, 238, 239, 240] usually ***excludes*** colour transformation. Nonetheless, we instead show that ***colour transformation is very useful for domain-generic person re-id***.

Table 4.7: Universal learning *vs.* unsupervised learning.

Method	Market-1501				Duke				MSMT17				CUHK03				VIPeR		
Metric(%)	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
LOMO [42]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8	-	-	-	-	0.6	1.9	3.6	0.7	-	-	-
BOW [43]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3	-	-	-	-	2.1	4.6	7.0	1.9	-	-	-
ISR [116]	40.3	-	-	14.3	-	-	-	-	-	-	-	-	-	-	-	-	27.0	49.8	61.2
Dic [115]	50.2	-	-	22.7	-	-	-	-	-	-	-	-	-	-	-	-	29.6	54.8	64.8
TAUDL [56]	63.7	-	-	**41.2**	**61.7**	-	-	**43.5**	28.4	-	-	12.5	-	-	-	-	-	-	-
BUC [117]	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5	-	-	-	-	-	-	-	-	-	-	-
UML(MSMT17)	**68.2**	**83.1**	**87.6**	37.0	60.9	**83.0**	**87.5**	37.0	-	-	-	-	**12.6**	**25.4**	**33.4**	**13.4**	**36.4**	**57.9**	**67.4**
UML(Duke)	66.1	81.6	86.3	35.5	-	-	-	-	**35.5**	**48.2**	**53.7**	**12.2**	10.7	21.9	27.5	10.5	35.4	54.1	62.3
Supervised Learning	90.4	96.5	97.8	73.5	81.5	90.8	93.1	65.9	73.3	84.8	88.1	44.1	40.8	62.7	73.6	40.4	39.2	65.8	77.5

Table 4.8: Dataset statistics and evaluation setting.

Dataset	Train		Test	
	# ID	# Image	# ID	# Image
VIPeR [158]	316	632	316	632
CUHK03 [40]	767	7,368	700	6,728
Market-1501 [43]	751	12,936	750	19,732
DukeMTMC [48]	702	16,522	702	18,889
MSMT17 [41]	1,041	32,621	3,060	93,820

The closest works to our method are image synthesis based unsupervised domain adaptation re-id models [151, 152, 50, 48, 49, 153]. All of them aim to transfer the labelled source person identity information to unlabelled target domains. However, these existing methods are domain-specific, and often need a complex model training for every single target domain. This "train once, run once" strategy is less usable and more costly to real-world system development. In contrast, our method needs neither domain-specific training nor difficult model optimisation (such as GANs [156]). We take a "train once, run everywhere" strategy based on simple and domain-generic image transformations. Our method uses flexibly off-the-shelf supervised learning re-id methods therefore can benefit continuously from a wide range of increasingly advanced learning algorithms developed by the wider community.

Table 4.9: Universal learning *vs.* unsupervised domain adaptation. *: Using more labelled source training data. †: Using additionally person attribute labels. UML uses the source data as the source training samples for fair comparison.

Source→Target	Duke→Market				Market→Duke				CUHK03→Market				CUHK03→Duke			
Metric (%)	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
UMDL† [52]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3	-	-	-	-	-	-	-	-
PUL [232]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4	41.9	57.3	64.3	18.0	23.0	34.0	39.5	12.0
CAMEL* [155]	54.5	-	-	26.3	-	-	-	-	-	-	-	-	-	-	-	-
TJ-AIDL† [53]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0	-	-	-	-	-	-	-	-
MMFA† [54]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7	-	-	-	-	-	-	-	-
DECAMEL* [154]	60.2	-	-	32.4	-	-	-	-	-	-	-	-	-	-	-	
PTGAN [41]	38.6	-	66.1	-	27.4	-	50.7	-	31.5	-	60.2	-	17.6	-	38.5	-
PoseNorm [152]	-	-	-	-	29.9	-	51.6	15.8	-	-	-	-	-	-	-	-
SPGAN [49]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3	42.3	-	-	19.0	-	-	-	-
SyRI* [151]	65.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CR-GAN[153]	59.6	-	-	29.6	**52.2**	-	-	**30.0**	58.5	75.8	81.9	30.4	**46.5**	**61.6**	**67.0**	**26.9**
HHL [50]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2	56.8	74.7	81.4	29.8	42.7	57.5	64.2	23.1
UML(Source)	**66.1**	**81.6**	**86.3**	**35.5**	46.3	**61.2**	**66.8**	26.7	**58.7**	**76.5**	**82.6**	**31.1**	42.8	57.8	64.3	23.2
Supervised Learning	90.4	96.5	97.8	73.5	81.5	90.8	93.1	65.9	90.4	96.5	97.8	73.5	81.5	90.8	93.1	65.9

4.2.2 Experiment

Datasets. We used five popular person re-id benchmarks in the standard train/test evaluation protocols. The statistics of these datasets are summarised in Table 4.8.

Performance metrics. We adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the model performance measurements.

Implementation details. To train a universal re-id model, we pre-trained the model on ImageNet and used the SGD algorithm with the momentum set to 0.9, the weight decay to 0.0005, and the mini-batch size of 32. We trained the model for totally 60 epochs, with the learning rate of 0.001 in the first 40 epochs, and the decay learning rate as 10 in the last 20 epochs. All input images were resized to 384×128 in pixel and subtracted by the ImageNet mean. On top of the proposed transformation strategy, we applied random cropping and flipping during training.

Universal Learning vs. *Unsupervised Learning*

We compared our UML with two hand-crafted feature models, (LOMO [42], BoW [43]), two dictionary learning models (ISR [116], Dic [115]), and two unsupervised deep learning methods (TAUDL [56], BUC [117]). In this test, we used MSMT17 and DukeMTMC-reID as the source training data, individually. Table 4.7 compares the performance of these methods. We have the following findings.

(1) Hand-crafted feature methods [42, 43] give the worst performance. This is due to weak representation without the ability to extract data relevant features. Also, they cannot optimise the matching metrics.

(2) Dictionary learning methods [116, 115] improve the performance by using reconstruction loss functions. However, their capability is limited by the input hand-crafted feature representations.

(3) More recent unsupervised learning models [56, 117] further push the performance envelope. In addition to per-domain model training requirement, these methods often come with some extra model parameters which are likely to be data sensitive. Typically, careful parameter tuning and costly model training are required in order to achieve competitive results. This is not favourable particularly for unsupervised learning where *no* labelled validation data available for hyper-parameter cross-validation and optimisation for the target domain.

(4) The proposed UML method matches or surpasses the performance of best competitors [56, 117] *without* training the model to the target domains. This suggests stronger domain generalisation and practical advantages of our method for the industrial adoption due to the "train once, run everywhere" merit. In reality, model training is costly in both budget and time. This therefore suggests an economical advantages and deployment-friendly superiority of our method over the strongest competitors in practice.

(5) Using MSMT17 as source leads to slightly better results as compared to using DukeMTMC. This is reasonable since MSMT17 offers more identities and images.

(6) Compared to supervised learning, unsupervised and universal learning models are clearly outperformed. This indicates a large room for further algorithm innovation.

Universal Learning vs. *Domain Adaptation*

We compared UML with state-of-the-art unsupervised domain adaptation re-id models, including 6 image synthesis models (PTGAN [41], PoseNorm [152], SyRI [151], SPGAN [49], HHL [50],

Table 4.10: Universal learning *vs.* hybrid strategy.

Method	Market-1501				MSMT17			
Metric (%)	R1	R5	R10	mAP	R1	R5	R10	mAP
ECN [57]	75.1	87.6	91.6	43.0	30.2	41.5	46.8	10.2
CASCL[157]	64.7	80.2	85.6	35.6	-	-	-	-
CR-GAN+TAUDL[153]	77.7	89.7	92.7	54.0	-	-	-	-
UML+TAUDL	**78.5**	**89.9**	**92.9**	**55.7**	**38.1**	**52.6**	**55.7**	**14.5**

CR-GAN [153]) and 6 feature alignment models (UMDL [52] CAMEL [155], PUL [232], TJ-AIDL [53], MMFA [54], DECAMEL [154]).

Table 4.9 shows that:

(1) UML is the best performer among all the competitors. Importantly, unlike previous approaches for *domain-specific* model training, our method needs only one time of *domain-agnostic* model training. Moreover, we do not require access the large number of unlabelled target data. This uniquely enables universal person re-id deployments.

(2) Feature alignment methods have obtained increasingly higher performance. It is worth noting that most feature learning methods such as DECAMEL unfairly benefit from extra labelled source data.

(3) Compared to feature alignment, image synthesis methods have started to achieve relatively superior cross-domain re-id accuracy. It is especially so considering that less label supervision is used (except SyRI).

(4) Both unsupervised domain adaptation and universal learning are significantly outperformed by the less scalable supervised learning, suggesting the necessity of devoting further more research efforts and endeavour for scaling state-of-the-art re-id methods.

Universal Learning vs. *Hybrid Strategy*

We compared UML with three state-of-the-art hybrid methods that combine unsupervised domain adaptation and unsupervised learning: ECN [57], CASCL [157], and CR-GAN [153] + TAUDL [56]. For a fair comparison, we combine UML with TAUDL [56] to exploit unlabelled target data. We tested the setting of Duke→Market/MSMT17. Table 4.10 shows UML+TAUDL is the best performer, suggesting the efficacy of our method in a hybrid learning scenario.

Table 4.11: Effect of Domain-generic image transformations (DIT). (source domain: Duke)

Source	Market				Duke				MSMT17				CUHK03				VIPeR		
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Duke	57.0	73.9	80.2	30.1	**83.5**	**91.8**	**94.3**	**70.0**	24.2	34.9	40.3	8.1	8.4	18.2	25.1	8.6	28.2	46.8	57.6
Duke+DIT	**66.1**	**81.6**	**86.3**	**35.5**	81.5	90.8	93.1	65.9	**35.5**	**48.2**	**53.7**	**12.2**	**10.7**	**21.9**	**27.5**	**10.5**	**35.4**	**54.1**	**62.3**
Gain(absolute)	+9.1	+7.7	+6.1	+5.4	-2.0	-1.0	-1.2	-4.1	+11.3	+13.3	+13.4	+4.1	+2.3	+3.7	+2.4	+1.9	+7.2	+7.3	+4.7
Gain(relative)	+16.0	+10.4	+7.6	+17.9	-2.4	-1.1	-1.3	-5.9	+46.7	+38.1	+33.3	+50.6	+27.4	+20.3	+9.6	+22.1	+25.5	+15.6	+8.2

Table 4.12: Effect of individual image transformations: Hue, Saturation, Lightness, and Contrast. (source domain: Duke)

Source	Market				Duke				MSMT17				CUHK03				VIPeR		
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Duke	57.0	73.9	80.2	30.1	**83.5**	**91.8**	**94.3**	**70.0**	24.2	34.9	40.3	8.1	8.4	18.2	25.1	8.6	28.2	46.8	57.6
Duke+**H**	62.5	78.4	83.6	32.2	82.0	91.2	94.0	68.2	26.3	37.3	42.7	8.6	8.6	17.8	22.9	8.3	30.7	48.7	57.0
Duke+**S**	60.3	76.8	82.4	31.8	82.3	91.2	94.2	68.7	26.5	37.6	42.7	8.9	8.8	18.8	25.6	9.0	29.4	46.2	55.1
Duke+**L**	59.6	75.2	80.9	31.1	82.4	91.4	94.1	68.5	27.1	38.2	43.5	9.2	10.6	20.6	26.6	10.4	34.5	53.2	59.5
Duke+**C**	59.4	74.3	81.1	30.7	83.4	91.6	94.1	69.0	29.8	41.7	47.3	10.1	10.1	20.0	26.0	10.1	33.5	52.2	62.7
Duke+**All**	**66.1**	**81.6**	**86.3**	**35.5**	81.5	90.8	93.1	65.9	**35.5**	**48.2**	**53.7**	**12.2**	**10.7**	**21.9**	**27.5**	**10.5**	**35.4**	**54.1**	**62.3**

Further Analysis and Discussions

Effect of domain-generic image transformations. We evaluated the effect of the proposed domain-generic image transformations. To this end, we compared the results without using our transformations on training data. We tested DukeMTMC-reID as the source domain. Table 4.11 shows that the proposed image transformation is consistently beneficial for improving the model performance on diverse target domains with very different camera viewing conditions. Both the absolute and relative performance gains are significant in most cases. As we aim for a domain-generic universal person re-id, the performance may be inferior to domain-specific models. To examine this, we compared UML with the supervised learning model (see the part with grey background). We indeed observe a performance drop but importantly *insignificant*, as humans tend to forget some old knowledge marginally whilst acquiring new knowledge over time. This means that our model can be similarly effective for the source domain as the supervised learning method. In contrast, image synthesis methods often suffer dramatic performance degradation on the source domain after adaptation, *i.e.*, the notorious *catastrophic forgetting* problem [241]. For example, SPGAN [49] experiences a Rank-1 drop of 16.4% (83.5%-67.1%) on the source

Table 4.13: SPGAN on top of UML.

Source	Method	Market			
		R1	R5	R10	mAP
Duke	UML	**66.1**	**81.6**	**86.3**	**35.5**
	UML+SPGAN	65.5	81.2	85.8	35.2

DukeMTMC-reID.

Types of image transformations. We examined the contribution of every individual image transformation: Hue, Saturation, Lightness, and Contrast. Table 4.12 shows that the performance benefits by individual transformations vary with test domains. This is reasonable due to the difference in the viewing condition characteristics of distinct domains which typically present no regularity. This also indicates the necessity of exploiting all the image transformations for tackling the domain heterogeneity during deployment at scale.

Domain universality. We quantified how well the UML model is generic and universal to various domains in the sense of being robust to transformations. We used the UML model trained with DukeMTMC-reID as the source domain, and tested its universality degree on transformed Market-1501 images. We selected randomly 1,000 Market-1501 source (original) images and applied individual transformations to each of them. Composited transformations were not used for simplified and dedicated analysis. Such transformations imitate the cross-domain person appearance variations. As a comparison, we tested a *baseline* model trained without using our image transformations.

We considered two measures of domain universality: (1) *Feature level* invariance, and (2) *Prediction level* invariance. The former is obtained by computing the Euclidean distance of the features of the transformed images against that of the original images all extracted by the UML model. The latter instead is quantified by the Euclidean distance between their classification prediction vectors. Figure 4.9 shows that UML enables to learn significantly invariant features w.r.t. image transformations therefore more robust re-id deployment across heterogeneous domains. This is consistent with the observations made in Tables 4.9 and 4.11.

Comparison to state-of-the-art image synthesis. As an image generation method, we specially compared our UML with the state-of-the-art image synthesis model SPGAN [49]. We did not select HHL [50] since it is a hybrid of image synthesis and cross-domain feature learning. We have

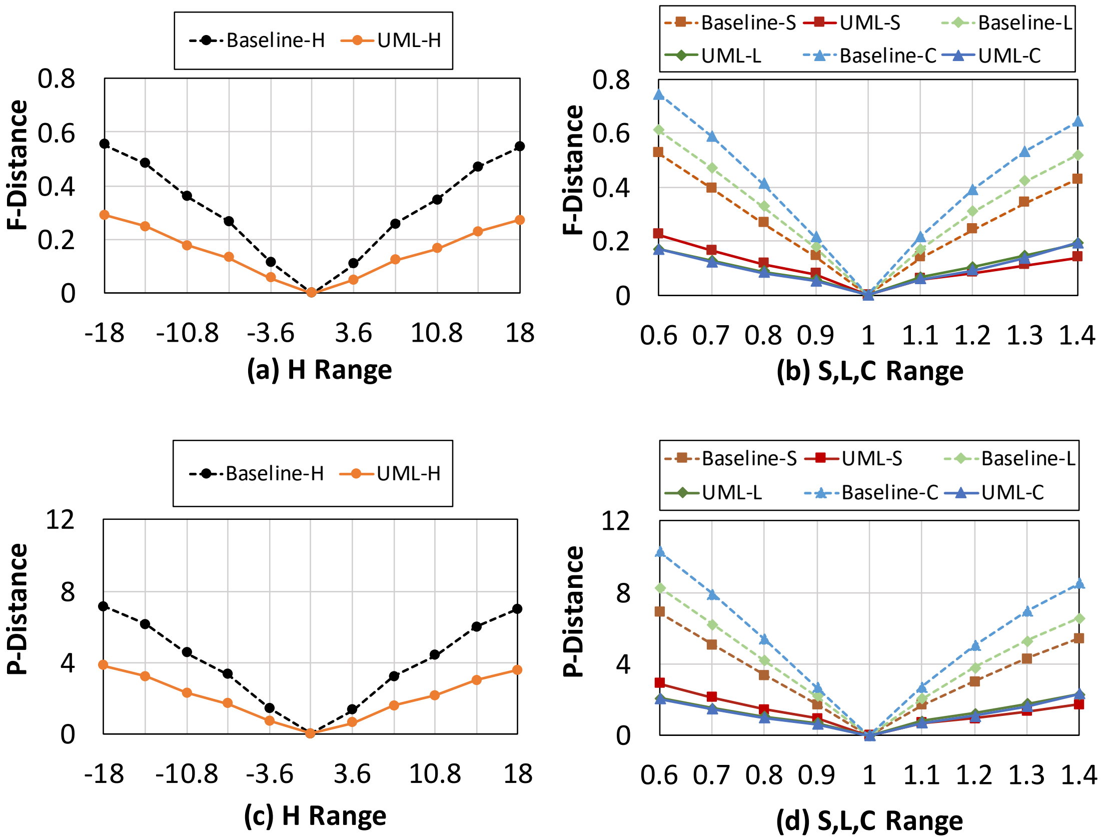


Figure 4.9: Domain universality analysis. source/test domain: DukeMTMC/Market. F=Feature, P=Prediction, H=Hue, S=Saturation, L=Lightness, C=Contrast.

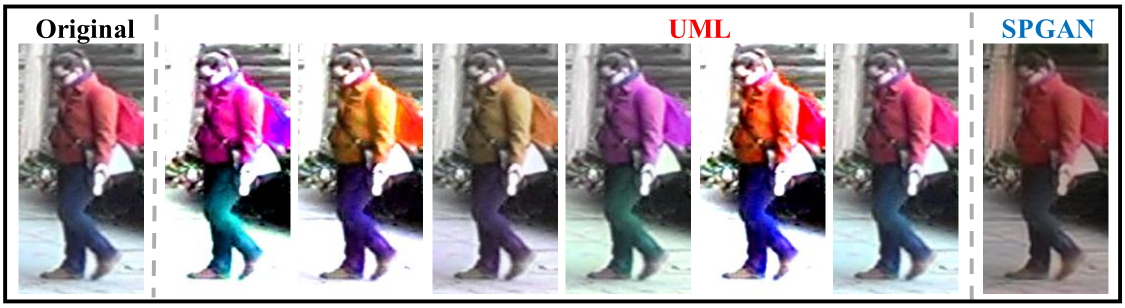


Figure 4.10: Visual comparison on DukeMTMC-reID.

already compared the quantitative results in Table 4.9, and showed that SPGAN is inferior. This is intuitively reasonable as observed from their visual comparison in Figure 4.10. Specifically, UML generates much more diverse images than SPGAN in a computationally more efficient and domain generic manner. On the contrary, SPGAN requires computationally expensive domain-specific model training along with tedious hyper-parameter tuning. By only altering the colour and contrast properties, UML can well preserve the person identity class information without the need for designing identity preserving loss function. The colour of clothing and/or associations may clearly changes w.r.t. the original source images, but all other identity information including person physical and biometric characteristics remain. This is partly against the conventional understanding that clothing colour plays the dominant role in person re-id therefore their variation of the same person identity class may hurt the model generalisation [6]. Our investigation and finding uniquely challenge this classical wisdom and validate the importance of otherwise appearance information to person re-id. This inspires future novel ideas especially for image synthesis modelling. Functionally, SPGAN images can be considered as part of UML images. To demonstrate this, we tested the complementary effect of SPGAN on top of UML. The results in Table 4.13 show that very limited effect can be resulted from adding SPGAN images to the training set. This also justifies the superior performance of UML since acquisition of large scale training data is one of the key elements for better model generalisation capability.

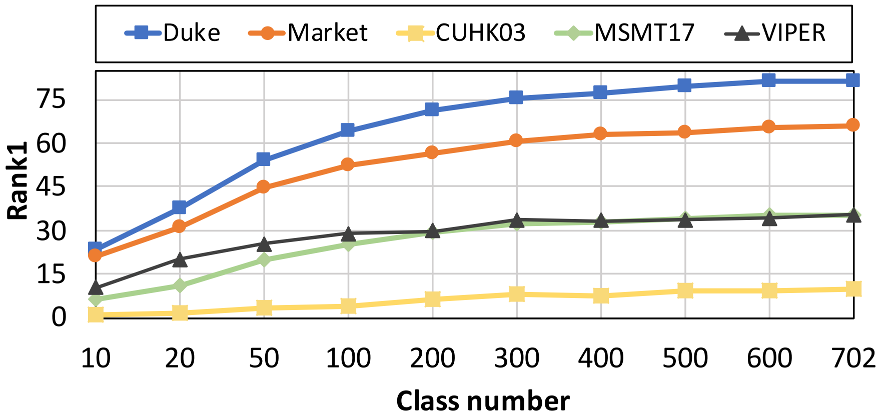


Figure 4.11: Effect of source (DukeMTMC) identity number.

source identity number. We tested the effect of source identity number on model performance. We used DukeMTMC as source training set and varied the identity number between 100 and 702. Figure 4.11 shows that more source classes generally lead to better performance as expected.

Surprisingly, our method is able to perform well using as few as 100 person classes ($\frac{1}{7}$ of the standard training size). This validates the efficacy of our model in case of limited source data.

4.3 Summary

This chapter discusses the cross-domain knowledge transfer for person re-identification. Section 4.1 presents a novel Hierarchical Unsupervised Domain Adaptation (HUDA) person re-id model for more discriminative domain adaptation from a labelled source domain to an unlabelled target domain. HUDA is designed for simultaneous global distribution alignment and local instance alignment. It addresses the limitations of existing unsupervised domain adaptation re-id models where only global distribution alignment is considered. Extensive evaluations validate the advantages of HUDA over state-of-the-art models. Besides, this chapter discusses a novel universal re-id problems. Section 4.2 presents a *Universal Model Learning* (UML) for domain-generic universal person re-id in a "train once, run everywhere" pattern. This differs from all the existing state-of-the-art supervised and unsupervised learning (including domain adaptation) methods typically taking a "train once, run once" pattern, suffering from per-domain *repeated* model training as well as the corresponding various costs and limitations. Our method therefore opens up a direction taking intelligent learning algorithms closer to industrial-level applications, although the current performance achieved is still inferior to that of supervised learning counterparts. As a training image generation method, our method is readily able to integrate any off-the-shelf supervised learning algorithms without extra complexity and obstacle of hyper-parameter tuning and model optimisation as required by image synthesis methods. We have conducted extensively comparative experiments for unsupervised person re-id in the unlabelled target domain using five public benchmarks, and demonstrated the performance superiority and modelling advantages of UML over the state-of-the-art alternative methods in both unsupervised model learning and unsupervised domain adaptation settings.

Chapter 5

Knowledge Transfer Across Models in Image Classification

If you have knowledge, let others light their candles in it.

—— **Margaret Fuller**

This chapter focuses on cross-model knowledge transfer, in particular, knowledge transfer on image classification is studied. To alleviate the expensive cost of the teacher model training and complex multi-stages scheme in knowledge transfer, this chapter investigates two different techniques: (1) Self-Referenced Deep Training (SRDT) and (2) Knowledge Distillation by On-the-fly Native Ensemble (ONE) approaches. The former focuses on resource-limited scenarios, while the latter performs well on reducing the complex training stages in cross-model knowledge transfer.

5.1 Self-Referenced Deep Training

In this work, we formulate a novel deep learning approach that improves the model generalisation capability through employing self-discovered knowledge as additional supervision signal with marginal extra computational cost and hence not hurting the computing scalability. We call this strategy **Self-Referenced Deep Learning** (SRDL), in contrast to Vanilla Deep Model training in Section 2.3.1.

5.1.1 Self-Referenced Deep Learning

SRDL Overview. The proposed SRDL approach is a knowledge referenced end-to-end deep model training strategy. The overview of our SRDL approach is depicted in Figure 5.1. This is realised through reformulating the vanilla training process into two equal-sized stages:

1. In the first stage (Figure 5.1(i)), SRDL learns the target model as a vanilla algorithm with a conventional supervised learning objective, while tries to induce reliable knowledge.

2. In the second stage (Figure 5.1(ii)), SRDL continues to train the model by a conventional supervised loss and a self-discovered knowledge guided imitation loss concurrently.

For model training, SRDL consumes the same number of epochs as the vanilla counterpart. The extra marginal cost is due to self-discovered knowledge extraction (see `Evaluation Metrics` in Sec 5.1.2). Consequently, SRDL allows to benefit model generalisation as knowledge distillation at faster optimisation speed. Once the target model is trained, it is deployed to the test data same as the vanilla method.

(I) First Stage Learning. In the first stage of SRDL, we train the deep model $\boldsymbol{\theta}$ by the cross-entropy loss Eq (2.3). Model training is often guided by a learning rate decay schedule such as the step-decay function [5, 8]:

$$\varepsilon_t = \varepsilon_0 \times f_{\text{step}}(t, M), \ \ t \in [1, \cdots, M] \tag{5.1}$$

where ε_t denotes the learning rate at the t-th epoch (initialised as ε_0, in total M epochs), and $f_{\text{step}}(t, M)$ the step-decay function. The learning rate decay aims to encourage the model to converge to a satisfactory local minimum without random oscillation in loss reduction during model training. However, if applying the conventional step-decay scheme throughout the optimisation process, SRDL may result in premature knowledge during training. This is because, the model still resides in an unstable local minimum due to that the learning rate drop is not sufficiently quick [242].

To overcome this problem, we propose to deploy an *individual* and *complete* step-decay schedule for both first and second stages of SRDL (Figure 5.1(c)), subject to the condition of remaining the same training epochs (cost). Formally, this schedule is expressed as:

$$\varepsilon_t = \varepsilon_0 \times f_{\text{step}}(t, 0.5M) \tag{5.2}$$

The intuition is that, the in-training model can be *temporarily* pushed towards a reasonably stable local minimum within the same number of (e.g. $0.5M$) epochs to achieve a more-ready state

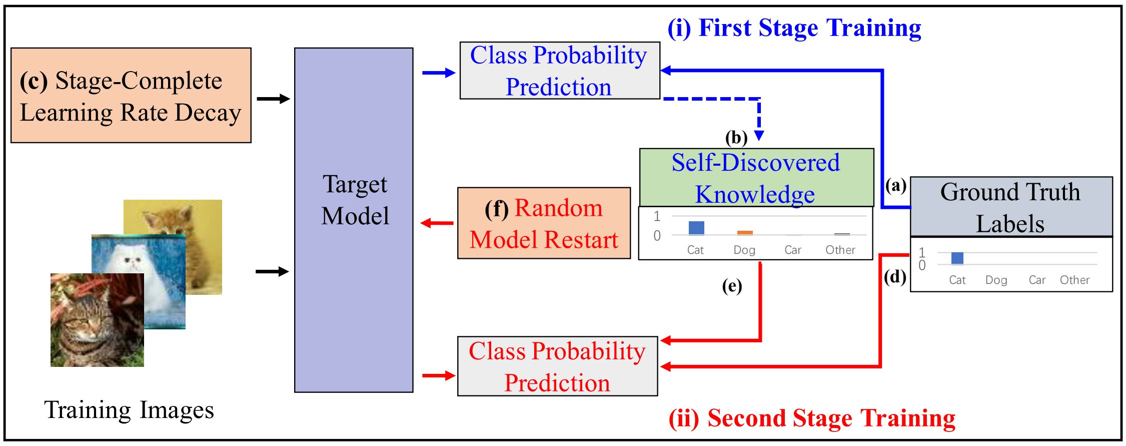

(i) First Stage Training
(c) Stage-Complete Learning Rate Decay
Target Model
Class Probability Prediction
(b)
Self-Discovered Knowledge
(a)
Ground Truth Labels
(f) Random Model Restart
1
0
Cat
Dog
Car
Other
(d)
(e)
Class Probability Prediction
Training Images
(ii) Second Stage Training

Figure 5.1: Overview of the proposed Self-Referenced Deep Learning (SRDL). The SRDL strategy consists of two stages training: **First stage**: We train the target model by a cross-entropy loss (Eq (2.3)) with **(a)** the available label supervision for half epochs, whilst learning to **(b)** extract discriminative intermediate knowledge concurrently (Eq (2.2)). To maximise the quality of self-discovered knowledge, we introduce **(c)** a pass-complete learning rate decay schedule (Eq (5.2)). **Second stage**: we continuously optimise the target model for the other half epochs by the joint supervision (Eq (5.3)) of both **(d)** the label data and **(e)** self-discovered intermediate knowledge in an end-to-end manner. We **(f)** randomly restart the model for the second stage to break the optimisation search space constraint from self-referenced deep learning mechanism.

therefore help ensure the quality of self-discovered knowledge. We call this a ***stage-complete*** learning rate step-decay schedule (Figure 5.2). Our evaluations verify the significance of this design while guaranteeing the goodness of the self-referenced knowledge (see Table 5.4).

At the end of the first stage of SRDL with a “half-trained” model (denoted as $\boldsymbol{\theta}^*$), we extract the self-discovered knowledge in the form of per-sample class probability prediction (Figure 5.1(b)). Formally, we compute the class probability $\tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta})$ for each training sample $\boldsymbol{x}$ by a *softened* softmax operation as Eq (2.2): We set T=3 in our experiments as suggested in [25].

(II) Second Stage Learning. To improve the generalisation performance of the model, we use the self-discovered knowledge to provide training experience at second stage model learning in SRDL. We quantify the imitation of the current model softened class prediction $\tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta})$ to the knowledge $\tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta}^*)$ with Kullback Leibler (KL) divergence (Figure 5.1(e)) by Eq (2.4).

The overall loss function for the second stage in SRDL is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + T^2 * R_{\text{kl}} \tag{5.3}$$

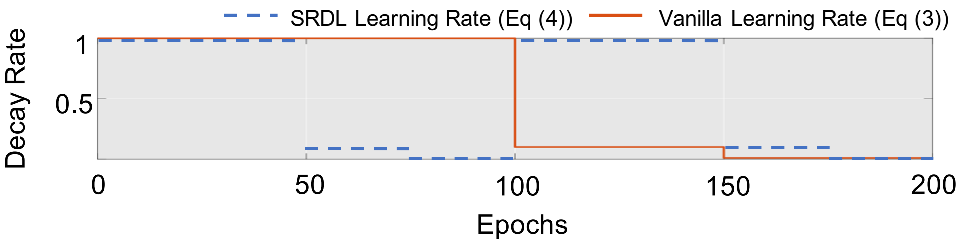


Figure 5.2: Illustration of a vanilla learning rate step-decay function and the proposed stage-complete learning rate step-decay schedule.

with the squared softening-temperature (Eq (2.2)) as the balance weight. The gradient magnitudes produced by the soft targets $\tilde{p}$ are scaled by $\frac{1}{T^2}$, so we multiply the distillation loss term by a factor T^2 to ensure that the relative contributions of ground-truth and teacher probability distributions remains. In doing so, the network model learns to both predict the correct class label (cross-entropy loss $\mathcal{L}_{\text{ce}}$) and align the class probability of previous training experience (imitation loss R_{kl}) concurrently.

Random Model Restart. A key difference between SRDL and knowledge distillation is that SRDL enables a model to learn from its own (previously revealed) knowledge through training experience rather than from an independent teacher's knowledge. This self-discovered knowledge is represented in the "half-trained" model parameters $\boldsymbol{\theta}^*$. If we further train the model at the second stage from $\boldsymbol{\theta}^*$ by Eq (5.3), the learning may become less explorable for better local or global minimum due to the stacking effect of the imitation loss and the model parameter status. Therefore, we start the second stage training with *randomly* initialised model parameters.

This scheme is based on three considerations: (1) A large proportion of the knowledge learned in the first stage has been extracted and used in the second stage. (2) The same training data will be used. (3) Random initialisation offers another opportunity for the model to converge to a better local minimum. Our experiment validates the effectiveness of this *random restart* scheme (see Table 5.5 in Sec 5.1.2).

SRDL model training is summarised in Alg 1. In our experiments, a SRDL trained model is tested against both the vanilla model training strategy and the knowledge distillation method.

Algorithm 1 Self-Referenced Deep Learning

```
Input: Labelled training data $\mathcal{D}$; Training epochs $M$;
Output: Trained CNN model $\boldsymbol{\theta}$;
(I) First stage learning
Initialisation: t=1; Random model $\boldsymbol{\theta}$ initialisation;
while $t \leq 0.5 * M$ do
    (i) Update the learning rate $\varepsilon_t$ (Eq (5.2));
    (ii) Update $\boldsymbol{\theta}$ by cross-entropy loss (Eq (2.3));
end
Knowledge Extraction Induce per-sample class probability predictions (Eq (2.2));
(II) Second stage learning
Initialisation: t=1; Random model $\boldsymbol{\theta}$ restart;
while $t \leq 0.5 * M$ do
    (i) Update the learning rate $\varepsilon_t$ (Eq (5.2));
    (ii) Update $\boldsymbol{\theta}$ by soft-feedback referenced loss (Eq (5.3));
end
```

5.1.2 Experiments

Experimental Setup

Datasets. For experimental evaluations, we use four benchmarking datasets including both coarse-grained object classification and fine-grained person instance identification Specifically, the ***CIFAR*10** and ***CIFAR*100** [81] datasets contain 32×32 sized natural images from 10 and 100 object classes. Both adopt a 50,000/10,000 train/test image split. The ***Tiny ImageNet*** [82] consists of 110,000 64×64 images from 200 object classes. We adopt the standard 100,000/ 1,000 train/val setting. The ***ImageNet [16]*** is a large scale 1,000-class object image classification benchmark, providing 1.2 million images for training, and 50,000 images for validation. The ***Market-1501*** [43] is a person re-identification dataset. Different from image classification as tested in the above four datasets, person re-identification is a more fine-grained recognition problem of matching person instance across non-overlapping camera views. It is a more challenging task due to the inherent zero-shot learning knowledge transfer from seen classes (identities) to unseen classes in deployments, i.e. no overlap between training and test classes. Market-1501 has 32,668 images of 1,501 different identities (ID) captured by six outdoor cameras. We use the standard 751/750 train/test ID split. Following [243, 20], we train the network by the cross-entropy loss (Eq (2.3)) and use the feature layer's output as the representation of person bounding box images for test by the Euclidean distance metric.

Performance Metrics. For performance measurement, we adopt the top-1 classification accuracy for image classification, the standard Cumulative Matching Characteristic (CMC) accuracy (Rank-n rates) and mean Average Precision (mAP) for person instance recognition (re-id). The CMC is computed for each individual rank k as the cumulative percentage of the truth matches for probes returned at ranks $\leq k$. And the Rank-1 rate is often considered as the most important performance indicator of an algorithm's efficacy. The mAP is to measure the recall of multiple truth matches, computed by first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes. We measure the model optimisation complexity with the FLoating-point OPerations (FLOPs): `Forward-FLOPs` * `Epochs` * `Training-Set-Size`.

Neural Networks. We use 7 networks in our experiments: one typical student net, ResNet-32 [5]; two typical teacher nets, ResNet-110 [5] and Wide ResNet WRN-28-10 [13]; and four varying sized nets, ResNet-50, DenseNet-121, DenseNet-201 and DenseNet-BC (L=190, k=40) [8].

Implementation Details. For all three image classification datasets, we use SGD with Nesterov momentum and set the mini-batch size to 128, the initial learning rate to 0.1, the weight decay to 0.0002, and the momentum to 0.9. For Market-1501, we use the same SGD but with the mini-batch size of 32. We assign sufficient epochs to all models to ensure convergence. On CIFAR datasets, the training budget is 300 epochs for DenseNet, and 200 epochs for ResNet and Wide ResNet models, same as [244]. We set 150/120 epochs on Tiny ImageNet/Market-1501 for all models. All model optimisation methods take the same epochs to train the target networks. We adopt a common learning rate decay schedule [244]: the learning rate drops by 0.1 at the 50% and 75% epochs. The data augmentation includes horizontal flipping and randomly cropping from images padded by 4 pixels on each side with missing pixels filled by original image reflections [5]. We report the average performance of 5 independent runs for each experiment.

Comparison with the Vanilla Learning Strategy

We compared the image classification performance between SRDL and the vanilla optimisation strategy. We make the following observations from Table 5.1:

Table 5.1: Comparison between SRDL and the vanilla learning strategy on image classification. Metric: Accuracy (Acc) Rate (%). "Gain": the performance gain by SRDL over vanilla. TrCost: Model training cost in unit of 10^{16} FLOPs, **lower is better**. M: Million. The first/second best results are in **red**/**blue**.

Dataset	# Param	CIFAR10		CIFAR100		Tiny ImageNet	
Metrics		Acc	TrCost	Acc	TrCost	Acc	TrCost
ResNet-32+vanilla	0.5M	92.53	0.08	69.02	0.08	53.33	0.32
ResNet-32+**SRDL**		**93.12**	0.08	**71.63**	0.08	**55.53**	0.32
Gain (SRDL-vanilla)		+0.59	0	+2.61	0	+2.20	0
WRN-28-10+vanilla	36.5M	94.98	12.62	78.32	12.62	58.38	50.48
WRN-28-10+**SRDL**		**95.41**	12.62	**79.38**	12.62	**60.80**	50.48
Gain (SRDL-vanilla)		+0.43	0	+1.06	0	+2.42	0
DenseNet-BC+vanilla	25.6M	**96.68**	10.24	**82.83**	10.24	**62.88**	40.96
DenseNet-BC+**SRDL**		**96.87**	10.24	**83.59**	10.24	**64.19**	40.96
Gain (SRDL-vanilla)		+0.19	0	+0.76	0	+1.31	0

1. All three networks ResNet-32, WRN-28-10, and DenseNet-BC improve the classification performance when trained by the proposed SRDL. For example, ResNet-32 achieves an accuracy gain of 0.59% on CIFAR10, of 2.61% on CIFAR100, and of 2.20% on Tiny ImageNet. This suggests the applicability of SRDL to standard varying-capacity network architectures.

2. SRDL achieves superior model generalisation performance with nearly zero extra model training cost[1].

3. Smaller network (ResNet-32) with fewer parameters generally benefits more from SRDL in model generalisation performance, making our method more attractive to resource-limited applications. Hence, our SRDL addresses the notorious hard-to-train problem in small networks to some degree [26].

Results on ImageNet. We test the large scale ImageNet with DenseNet201 and obtain the

[1] The computational cost of knowledge extraction required by both SRDL and Knowledge Distillation [25] is marginal (less than 0.67% model training cost) and hence omitted for analysis convenience.

Top-1/5 rates 77.20%/94.57% by the vanilla vs 77.72%/94.89% by our SRDL. This suggests that SRDL generalises to large scale object classification settings.

Comparison with Knowledge Distillation

We compared our SRDL with the closely related Knowledge Distillation (KD) method [25]. With KD, we take ResNet-32 as the target model, WRN-28-10 and ResNet-110 as the pre-trained teacher models to produce the per-sample class probability targets (i.e. the teacher's knowledge) for the student. From Table 5.2 we draw these observations:

Table 5.2: Comparison between SRDL and Knowledge Distillation (KD) on image classification. Metric: Accuracy (Acc) Rate (%). TrCost: Model training cost in unit of 10^{16} FLOPs, **lower is better**. Number in bracket: model parameter size. The first/second best results are in **red**/**blue**.

Target Net	Method	Teacher Net	CIFAR10		CIFAR100		Tiny ImageNet	
			Acc	TrCost	Acc	TrCost	Acc	TrCost
ResNet-32 (0.5M)	Vanilla	N/A	92.53	**0.08**	69.02	**0.08**	53.33	**0.32**
	KD	WRN-28-10 (36.5M)	**92.83**	12.70	**72.58**	12.70	**56.80**	50.80
		ResNet-110 (1.7M)	92.75	0.30	71.17	0.30	55.06	1.20
	SRDL	N/A	**93.12**	**0.08**	**71.63**	**0.08**	**55.53**	**0.32**

1. KD is indeed effective to improve small model generalisation compared to the vanilla optimisation, particularly when using a more powerful teacher (WRN-28-10). However, this is at the price of extra 157× (12.70/0.08-1 or 50.80/0.32-1) model training cost. When using ResNet-110 as the teacher in KD, the performance gain is less significant.

2. SRDL approaches the performance of KD(WRN-28-10) on CIFAR100 and Tiny ImageNet, whilst surpasses it on CIFAR10. This implies that while small model is inferior to KD in self-discovering knowledge among a large number of classes, it seems to be superior for small scale tasks with fewer classes.

3. SRDL consistently outperforms KD(ResNet-110) in both model performance and training cost, indicating that KD is not necessarily superior than SRDL in enhancing small model generalisation (teacher dependent). This may be partly due to the overfitting of a stronger

teacher model (e.g. ResNet-110) which leads to less extra supervision information. To test this, we calculated the average cross entropy loss of the final epoch. We observed 0.0087 (ResNet-110) vs 0.1637 (ResNet-32), which is consistent with our hypothesis.

Table 5.3: Evaluation of person re-id (instance recognition) on Market-1501. The first/second best results are in **red**/**blue**.

Query Type	Single-Query		Multi-Query	
Metrics (%)	Rank-1	mAP	Rank-1	mAP
SCS [202]	51.9	26.3	-	-
G-SCNN [245]	65.8	39.5	76.0	48.4
HPN [246]	76.9	-	-	-
MSCAN [247]	80.3	57.5	86.8	66.7
JLML [20]	85.1	65.5	89.7	74.5
SVDNet [243]	82.3	62.1	-	-
PDC [248]	84.1	63.4	-	-
TriNet [31]	84.9	69.1	90.5	76.4
IDEAL [32]	86.7	67.5	91.3	76.2
DPFL [138]	88.6	72.6	92.2	80.4
BraidNet-CS+SRL [36]	83.7	69.5	-	-
DaRe [249]	86.4	69.3	-	-
ResNet-50+vanilla	87.5	69.9	91.4	78.5
ResNet-50+**SRDL**	**89.3**	**73.5**	**93.1**	**81.5**
Gain (SRDL-vanilla)	+1.8	+3.6	+1.7	+3.0
DenseNet-121+vanilla	**90.1**	74.0	**93.6**	81.7
DenseNet-121+**SRDL**	**91.7**	**76.8**	**94.2**	**83.5**
Gain (SRDL-vanilla)	+1.6	+2.8	+0.6	+1.8

Evaluation on Person Instance Recognition

In person re-identification (re-id) experiment, we compared SRDL with the vanilla model learning strategy using the same CNN nets, and also compared with ten recent the state-of-the-art re-id methods. Two different networks are tested: ResNet-50 (25.1M parameters) and DenseNet-121 (7.7M parameters). Table 5.3 shows that:

1. All CNN models benefit from SRDL on the person re-id task, boosting the re-id performance for both single-query and multi-query settings.

2. SRDL trained CNNs show superior re-id performance over most state-of-the-art methods. In particular, SRDL trained DenseNet-121 achieves the best re-id matching rates among all the competitors.

Note that, this performance gain is obtained from a general-purpose network without applying any specialised person re-id model training "bells and whistles". This is in strong contrast to existing deep re-id methods [245, 248, 247, 246] where specially designed network architectures with complex training process are required in order to achieve the reported results.

Component Analysis and Discussion

We further conducted SRDL component analysis using ResNet-32 on CIFAR100.

Stage-Complete Schedule. Table 5.4 compares our stage-complete learning rate decay schedule with the conventional *stage-incomplete* counterpart. It is evident that without the proposed schedule, self-referenced learning can be highly misleading due to unreliable knowledge extracted from the "half-trained" model. This validates the aforementioned model optimisation behaviour consideration (see the discussion underneath Eq (5.2)).

Table 5.4: Stage-complete schedule.

Decay Strategy	Accuracy (%)
Stage-Incomplete	58.11
Stage-Complete	**71.63**

Random Model Restart. Table 5.5 shows that model random restart for the second stage training in SRDL brings 1.90% (71.63%-69.73%) accuracy gain. This verifies our design motivation that the discriminative knowledge is well preserved in the training data and self-discovered correlation; Hence, random model initialisation for the second stage training of SRDL enables to break the optimisation search space constraint without losing the available information, and eventually improving the model generalisation capability.

Table 5.5: Random model restart.

Random Restart	Accuracy (%)
✗	69.73
✓	**71.63**

Model Ensemble. Table 5.10 shows that the ensemble of "half-trained" and final models can further boost the performance by 0.70% (72.33%-71.63%) with more (double) deployment cost. This suggests that the two models induced sequentially during training are partially complementary, which gives rise to model ensembling diversity and results in model performance boost. Besides, we also tested an ensemble of two randomly initialised networks each trained by the vanilla learning strategy for $M/2$ epochs, obtaining the Top-1 rate 72.02% vs 72.33% by SRDL. This shows that our SRDL ensemble outperforms the vanilla counterpart.

Table 5.6: Model ensemble.

Model Ensemble	Accuracy (%)
✗	71.63
✓	**72.33**

Model Generalisation Analysis. As shown in [250], model generalisation is concerned with the width of a local optimum. We thus examined the solutions $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_s$ discovered by the vanilla and SRDL training algorithms, respectively. We added small perturbations as $\boldsymbol{\theta}_*(d, \boldsymbol{v}) = \boldsymbol{\theta}_* + d \cdot \boldsymbol{v}$, $* \in \{v, s\}$ where $\boldsymbol{v}$ is a uniform distributed direction vector with a unit length, and $d \in [0, 5]$ controls the change magnitude. The loss is quantified by the cross-entropy measurement between the predicted and ground-truth labels. Figure 5.3 shows the robustness of each solution against the parameter perturbation, indicating the width of local optima as $\boldsymbol{\theta}_v < \boldsymbol{\theta}_s$. This suggests that our SRDL finds a wider local minimum than the vanilla therefore more likely to generalise better.

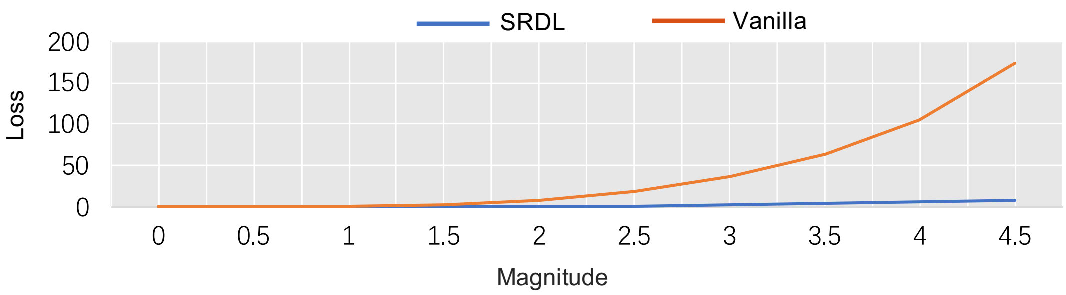


Figure 5.3: The width analysis of solution local optima.

5.2 Knowledge Distillation by On-the-Fly Native Ensemble

In this section, we propose a Knowledge Distillation by On-the-Fly Native Ensemble (ONE) model to improve the deep network optimization by online distillation. Compare to the proposed SRDL in Section 5.1, ONE is a one stage training procedure with a large capacity teacher network.

5.2.1 Methodology

We formulate an online distillation training method based on a concept of On-the-fly Native Ensemble (ONE). For understanding convenience, we take ResNet-110 [5] as an example. It is straightforward to apply ONE to other network architectures. For model training, we often have access to n labelled training samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_i^n$ with each belonging to one of C classes $y_i \in \mathcal{Y} = \{1, 2, \cdots, C\}$. The network $\boldsymbol{\theta}$ outputs a probabilistic class posterior $p(c|\boldsymbol{x}, \boldsymbol{\theta})$ for a sample $\boldsymbol{x}$ over a class c by Eq (2.1):

To train a multi-class classification model, we typically adopt the Cross-Entropy (CE) measurement $\mathcal{L}_{\text{ce}}$ (Eq (2.3)) between the predicted and ground-truth label distributions as the objective loss function. With the CE loss, the network is trained to predict the correct class label in a principle of maximum likelihood. To further enhance the model generalisation, we concurrently distil extra knowledge from an on-the-fly native ensemble (ONE) teacher to each branch in training.

On-the-Fly Native Ensemble. An overview of the ONE architecture is depicted in Figure 5.4. The ONE consists of two components: **(1)** m auxiliary branches with the same configuration (Res4X block and an individual classifier), each of which serves as an independent classifica-

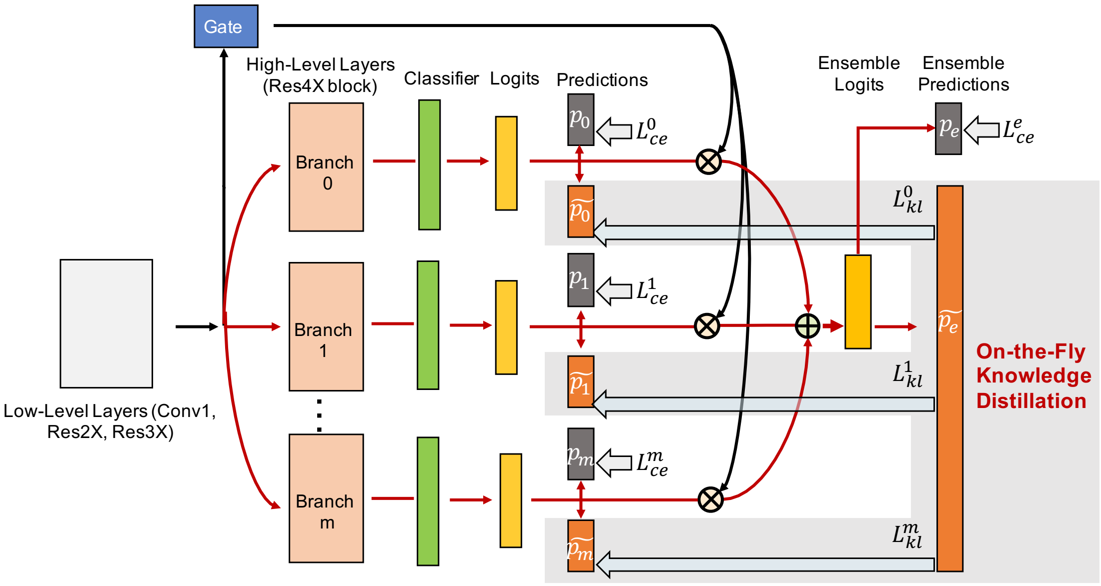


Figure 5.4: Overview of online distillation training of ResNet-110 by the proposed On-the-fly Native Ensemble (ONE). With ONE, we start by reconfiguring the target network by adding m auxiliary branches on shared low-level layers. All branches together with shared layers make individual models, all of which are then used to construct a stronger teacher model. During the mini-batch training process, we employ the teacher to assemble knowledge of branch models on-the-fly, which is in turn distilled back to all branches to enhance the model learning in a closed-loop form. In test, auxiliary branches are discarded or kept according to the deployment efficiency requirement.

tion model with shared low-level stages/layers. This is because low-level features are largely shared across different network instances and sharing them allows to reduce the training cost. **(2)** A gating component which learns to ensemble all $(m+1)$ branches to build a stronger teacher model. It is constructed by one fully connected (FC) layer followed by batch normalisation, ReLU activation, and softmax, using the same input features as the branches.

Our ONE method is established based on a multi-branch design specially for model training with several merits: (1) Enable the possibility of creating a strong teacher model without training a set of networks at a high computational cost; (2) Introduce a multi-branch simultaneous learning regularisation which benefits model generalisation (Figure 5.6(a)); (3) Avoid the tedious need for asynchronous update between multiple networks.

Under the reconfiguration of network, we add a separate CE loss $\mathcal{L}^i_{\text{ce}}$ to each branch which simultaneously learns to predict the same ground-truth class label of a training sample. While sharing the most layers, each branch can be considered as an independent multi-class classifier in that all of them independently learn high-level semantic representations. Consequently, taking the ensemble of all branches (classifiers) can make a stronger teacher model. One common way of ensembling models is to average individual predictions. This may ignore the diversity and importance variety of the member models of an ensemble. We therefore learn to ensemble by a gating component as:

$$\boldsymbol{z}_e = \sum_{i=0}^{m} g_i \cdot \boldsymbol{z}_i \tag{5.4}$$

where g_i is the importance score of the i-th branch's logits $\boldsymbol{z}_i$, and $\boldsymbol{z}_e$ is the logits of the ONE teacher. In particular, we denote the original branch as $i = 0$ for indexing convenience. We train the ONE teacher model with the CE loss $\mathcal{L}_{\text{ce}}^{e}$ (Eq (2.3)) the same as the branches.

Knowledge Distillation. Given the teacher's logits of each training sample, we distil this knowledge back into all branches in a closed-loop form. For facilitating knowledge transfer, we compute soft probability distributions at a temperature of T for individual branches $\tilde{p}_i(c|\boldsymbol{x}, \boldsymbol{\theta}^i)$ and the ONE teacher $\tilde{p}_e(c|\boldsymbol{x}, \boldsymbol{\theta}^e)$ as Eq (2.2), where i denotes the branch index, $i=0, \cdots, m$, $\boldsymbol{\theta}^i$ and $\boldsymbol{\theta}^e$ the parameters of the branch and teacher models respectively. Higher values of T lead to more softened distributions.

To quantify the alignment between individual branches and the teacher in their predictions, we use the Kullback Leibler divergence (Eq (2.4)) from branches to the teacher written as:

$$\mathcal{L}_{\text{kl}} = \sum_{i=0}^{m} \sum_{j=1}^{C} \tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e) \log \frac{\tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e)}{\tilde{p}_i(j|\boldsymbol{x}, \boldsymbol{\theta}^i)}. \tag{5.5}$$

Overall Loss Function. We obtain the overall loss function for online distillation training by the proposed ONE as:

$$\mathcal{L} = \sum_{i=0}^{m} \mathcal{L}_{\text{ce}}^{i} + \mathcal{L}_{\text{ce}}^{e} + T^2 * \mathcal{L}_{\text{kl}} \tag{5.6}$$

where $\mathcal{L}_{\text{ce}}^{i}$ and $\mathcal{L}_{\text{ce}}^{e}$ are the conventional CE loss terms associated with the i-th branch and the ONE teacher, respectively. The gradient magnitudes produced by the soft targets $\tilde{p}$ are scaled by $\frac{1}{T^2}$, so we multiply the distillation loss term by a factor T^2 to ensure that the relative contributions of ground-truth and teacher probability distributions remain roughly unchanged. Note, the entire ONE objective function of ONE is *not* an ensemble learning since (1) these loss functions corresponding to the models with different roles, and (2) the conventional ensemble learning often takes *independent* training of member models.

Model Training and Deployment. The model optimisation and deployment details are summarised in Alg 2. Unlike the two-phase offline distillation training, the target network and the

Algorithm 2 Knowledge Distillation by On-the-Fly Native Ensemble

```
**Input**: Labelled training data $\mathcal{D}$; Training epoch number $\tau$; Auxiliary branch number $m$;
**Output**: Trained target CNN model $\boldsymbol{\theta}^0$, and auxiliary models $\{\boldsymbol{\theta}^i\}_{i=1}^m$;
**/* Training */**
**Initialisation**: t=1; Randomly initialise $\{\boldsymbol{\theta}^i\}_{i=0}^m$;
**while** $t \leq \tau$ **do**
    Compute predictions of all individual branches $\{p_i\}_{i=0}^m$ (Eq (2.1));
    Compute the teacher logits (Eq (5.4));
    Compute the soft targets of all the branch and teacher models (Eq (2.2));
    Distil knowledge from the teacher back to all the branch models (Eq (5.5));
    Compute the final ONE loss function (Eq (5.6));
    Update the model parameters $\{\boldsymbol{\theta}^i\}_{i=0}^m$ by a SGD algorithm.
**end**
**/* Testing */**
**Single model deployment:** Use $\boldsymbol{\theta}^0$;
**Ensemble deployment (ONE-E):** Use $\{\boldsymbol{\theta}^i\}_{i=0}^m$.
```

ONE teacher are trained simultaneously and collaboratively, with the knowledge distillation from the teacher to the target being conducted in each mini-batch and throughout the whole training procedure. Since there is one multi-branch network rather than multiple networks, we only need to carry out the same stochastic gradient descent through $(m+1)$ branches, and training the whole network until converging, as the standard single-model incremental batch-wise training. There is no complexity of asynchronously updating among different networks which is required in deep mutual learning [33].

Once the model is trained, we simply remove all the auxiliary branches and obtain the original network architecture for deployment. Hence, our ONE method does not increase the test-time cost. However, if there is less constraint on the computation budget and the model performance is more important, we can deploy it as an ensemble model with all trained branches, denoted as "ONE-E".

5.2.2 Experiments

Datasets. We used four multi-class categorisation benchmark datasets in our evaluations (Figure 5.5). Three of them, including *CIFAR10*, *CIFAR100*, *ImageNet* are detailed in Section 5.1.1. We provide an additional SVHN dataset for the evaluation. The Street View House Numbers (SVHN) dataset consists of 73,257/26,032 standard training/text images and an extra set of 531,131 training images. We used all the training data *without* using data augmentation as [10].

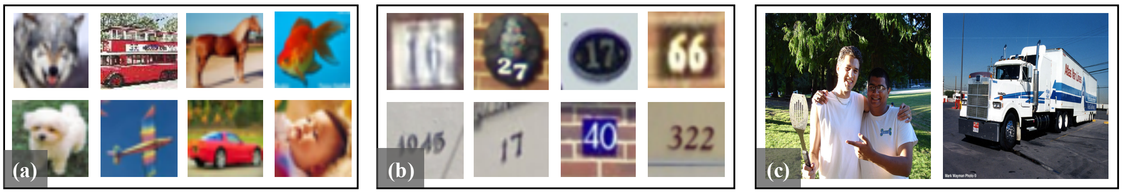


Figure 5.5: Example images from (a) CIFAR, (b) SVHN, and (c) ImageNet.

Performance Metrics. We adopted the common top-n (n=1, 5) classification error rate. To measure the computational cost of model training and test, we used the criterion of floating point operations (FLOPs). For any network trained by ONE, we reported the average performance of all branch outputs with standard deviation.

Experiment Setup. We implemented all networks and model training procedures in Pytorch. For all datasets, we adopted the same experimental settings as [251, 86] for making fair comparisons.

Table 5.7: Evaluation of our ONE method on CIFAR and SVHN. Metric: Error rate (%).

Method	CIFAR10	CIFAR100	SVHN	Params
ResNet-32 [5]	6.93	31.18	2.11	0.5M
ResNet-32 + **ONE**	**5.99±0.05**	**26.61±0.06**	**1.83±0.05**	0.5M
ResNet-110 [5]	5.56	25.33	2.00	1.7M
ResNet-110 + **ONE**	**5.17±0.07**	**21.62±0.26**	**1.76±0.07**	1.7M
ResNeXt-29(8×64d) [86]	3.69	17.77	1.83	34.4M
ResNeXt-29(8×64d) + **ONE**	**3.45±0.04**	**16.07±0.08**	**1.70±0.03**	34.4M
DenseNet-BC(L=190, k=40) [8]	3.32	17.53	1.73	25.6M
DenseNet-BC(L=190, k=40) + **ONE**	**3.13±0.07**	**16.35±0.05**	**1.63±0.05**	25.6M

We used the SGD with Nesterov momentum and set the momentum to 0.9. We deployed a standard learning rate schedule that drops from 0.1 to 0.01 at 50% training and to 0.001 at 75%. For the training budget, we set 300/40/90 epochs for CIFAR/SVHN/ImageNet, respectively. We adopted a 3-branch ONE ($m = 2$) design unless stated otherwise. We separated the last block of each backbone net from the parameter sharing (except on ImageNet we separated the last 2 blocks to give more learning capacity to branches) without extra structural optimisation (see ResNet-110 for example in Figure 5.4). Following [25], we set $T = 3$ in all the experiments. Cross-validation of this parameter T may give better performance but at the cost of extra model tuning.

Evaluation of On-the-Fly Native Ensemble

Results on CIFAR and SVHN. Table 5.7 compares top-1 error rate performances of four varying-capacity state-of-the-art network models trained by the conventional and our ONE learning algorithms. We have these observations: (1) All different networks benefit from the ONE training algorithm, particularly with small models achieving larger performance gains. This suggests a generic superiority of our method for online knowledge distillation from the on-the-fly teacher to the target student model. (2) All individual branches have similar performances, indicating that they have made sufficient agreement and exchanged respective knowledge to each other well through the proposed ONE teacher model during training.

Results on ImageNet. Table 5.8 shows the comparative performances on the 1000-classes ImageNet. It is shown that the proposed ONE learning algorithm again yields more effective

Table 5.8: Evaluation of our ONE method on ImageNet. Metric: Error rate (%).

Method	Top-1	Top-5
ResNet-18 [5]	30.48	10.98
ResNet-18 + **ONE**	**29.45±0.23**	**10.41±0.12**
ResNeXt-50 [86]	22.62	6.29
ResNeXt-50 + **ONE**	**21.85±0.07**	**5.90±0.05**
SeNet-ResNet-18 [252]	29.85	10.72
SeNet-ResNet-18 + **ONE**	**29.02±0.17**	**10.13±0.12**

Table 5.9: Comparison with knowledge distillation methods on CIFAR100. "*": Reported results. TrCost/TeCost: Training/test cost, in unit of 10^8 FLOPs. **Red**/**Blue**: Best and second best results.

Target Network	ResNet-32			ResNet-110		
Metric	Error (%)	TrCost	TeCost	Error (%)	TrCost	TeCost
KD [25]	**28.83**	6.43	1.38	N/A	N/A	N/A
DML [33]	29.03±0.22*	**2.76**	1.38	**24.10±0.72**	**10.10**	5.05
ONE	**26.61±0.06**	**2.28**	1.38	**21.62±0.26**	**8.29**	5.05

training and more generalisable models in comparison to the vanilla SGD. This indicates that our method is generically applicable in large scale image classification settings.

Comparison with Distillation Methods

We compared our ONE method with two representative distillation methods: Knowledge Distillation (KD) [25] and Deep Mutual Learning (DML) [33]. For the offline competitor KD, we used a large network ResNet-110 as the teacher and a small network ResNet-32 as the student. For the online methods DML and ONE, we evaluated their performances using either ResNet-32 or ResNet-110 as the target student model. We observed from Table 5.9 that: (1) ONE outperforms both KD (offline) and DML (online) distillation methods in error rate, validating the performance advantages of our method over alternative algorithms when applied to different CNN models. (2) ONE takes the least model training cost and the same test cost as others, therefore giving the most cost-effective solution.

Table 5.10: Comparison with ensembling methods on CIFAR100. "*": Reported results. TrCost/TeCost: Training/test cost, in unit of 10^8 FLOPs. **Red**/**Blue**: Best and second best results.

Network	ResNet-32			ResNet-110		
Metric	Error (%)	TrCost	TeCost	Error (%)	TrCost	TeCost
Snopshot Ensemble [244]	27.12	**1.38**	6.90	23.09*	**5.05**	25.25
2-Net Ensemble	26.75	2.76	**2.76**	22.47	10.10	**10.10**
3-Net Ensemble	**25.14**	4.14	4.14	**21.25**	15.15	15.15
ONE-E	**24.63**	**2.28**	**2.28**	**21.03**	**8.29**	**8.29**
ONE	26.61	2.28	1.38	21.62	8.29	5.05

Comparison with Ensembling Methods

Table 5.10 compares the performances of our multi-branch (3 branches) based model ONE-E and standard ensembling methods. It is shown that ONE-E yields not only the best test error but also enables most efficient deployment with the lowest test cost. These advantages are achieved at the second lowest training cost. Whilst Snapshot Ensemble takes the least training cost, its generalisation capability is unsatisfied with a notorious drawback of much higher deployment cost.

It is worth noting that ONE (without branch ensemble) already outperforms comprehensively a 2-Net Ensemble in terms of error rate, training and test cost. Comparing a 3-Net Ensemble, ONE approaches the generalisation capability whilst having larger model training and test efficiency advantages.

Table 5.11: Model component analysis on CIFAR100. Network: ResNet-110.

Configuration	Full	W/O Online Distillation	W/O Sharing Layers	W/O Gating
ONE	**21.62±0.26**	24.73±0.20	22.45±0.52	22.26±0.23
ONE-E	21.03	21.84	**20.57**	21.79

Model Component Analysis

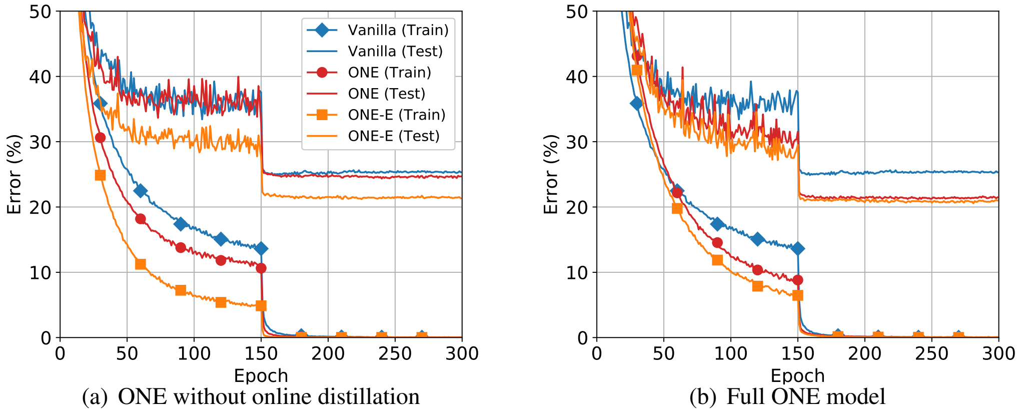
Table 5.11 shows the benefits of individual ONE components on CIFAR100 using ResNet-110. We have these observations: (1) **Without online distillation**, the target network suffers a performance drop of 3.11% (24.73-21.62) in test error rate. This performance drop validates the efficacy and quality of the ONE teacher in terms of performance superiority over individual branch models. This can be more clearly seen in Figure 5.6 that the ONE teacher fits better to

Table 5.12: Benefit of adding branches to ONE on CIFAR100. Network: ResNet-32.

Branch #	1	2	3	4	5
Error (%)	31.18	27.38	26.68	26.58	**26.52**

training data and generalises better to test data. Due to the closed-loop design, the ONE teacher also mutually benefits from distillation, reducing its error rate from 21.84% to 21.03%. With distillation, the target model effectively approaches the ONE teacher (Figure 5.6(a) vs 5.6(b)) on both training and test error performance, indicating the success of teacher knowledge transfer. Interestingly, even without distillation, ONE still achieves better generalisation than the vanilla algorithm. This suggests that our multi-branch design brings some positive regularisation effect by concurrently and jointly learning the shared low-level layers subject to more diverse high-level representation knowledge. (2) **Without sharing the low-level layers** not only increases the training cost (83% increase), but also leads to weaker performance (0.83% error rate increase). The plausible reason is a lack of multi-branch regularisation effect as indicated in Figure 5.6(a). (3) Using average ensemble of branches **without gating** (Eq (5.4)) causes a performance decrease of 0.64%(22.26-21.62). This suggests the benefit of adaptively exploiting the branch diversity in forming the ONE teacher.

The main experiments use 3 branches in ONE. Table 5.12 shows that ONE scales well with more branches and the ResNet-32 model generalisation improves on CIFAR100 with the number of branches added during training hence its performance advantage over the independently trained network (31.18% error rate).



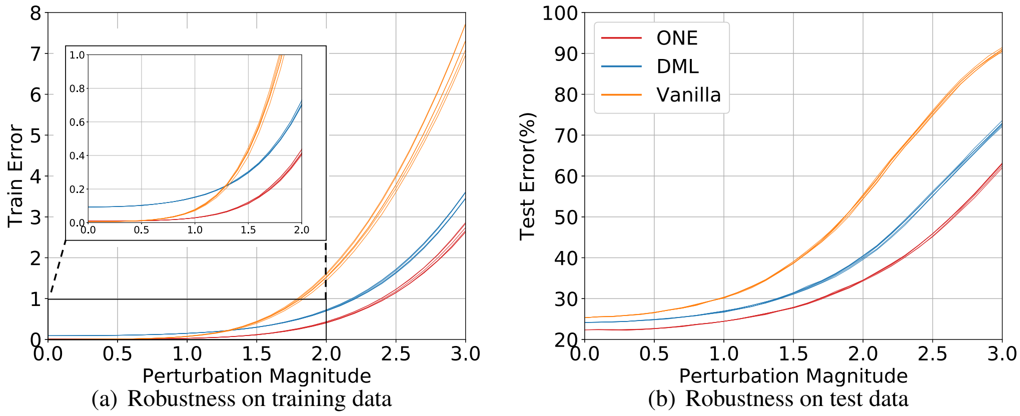
(a) ONE without online distillation

(b) Full ONE model

Figure 5.6: Effect of online distillation. Network: ResNet-110.

Model Generalisation Analysis

We aim to give insights on why an ONE trained network yields a better generalisation capability. A few previous studies [250] demonstrate that the width of a local optimum is related to the model generalisation. A general understanding is that, the surfaces of training and test error largely mirror to each other and it is favourable to converge the models to broader optima in training. As such, a trained model remains approximately optimal even under small perturbations at test time. Next, we exploited this criterion to examine the quality of model solutions $\boldsymbol{\theta}_v$, $\boldsymbol{\theta}_m$, $\boldsymbol{\theta}_o$ discovered by the vanilla, DML and ONE training algorithms respectively. This analysis was conducted on CIFAR100 using ResNet-110.



(a) Robustness on training data

(b) Robustness on test data

Figure 5.7: Robustness test of ResNet-110 solutions found by ONE, DML, and vanilla training algorithms on CIFAR100. Each curve corresponds to a specific perturbation direction $\boldsymbol{v}$.

Specifically, to test the width of local optimum, we added small perturbations to the solutions as $\boldsymbol{\theta}_*(d,\boldsymbol{v}) = \boldsymbol{\theta}_* + d \cdot \boldsymbol{v}$, $* \in \{v,m,o\}$ where $\boldsymbol{v}$ is a uniform distributed direction vector with a unit length, and $d \in [0,5]$ controls the change magnitude. At each magnitude scale, we further sampled randomly 5 different direction vectors to disturb the solutions. We then tested the robustness of all perturbed models in training and test error rates. The training error was quantified as the cross-entropy measurement between the predicted and ground-truth label distributions.

We observed in Figure 5.7 that: (1) The robustness of each solution against parameter perturbation appears to indicate the width of local optima as: $\boldsymbol{\theta}_v < \boldsymbol{\theta}_m < \boldsymbol{\theta}_o$. That is, ONE seems to find the widest local minimum among the three therefore more likely to generalise better than others.

(2) Comparing with DML, vanilla and ONE found deeper local optima with lower training

errors. This indicates that DML may probably get stuck in training, therefore scarifying the vanilla's exploring capability for more generalisable solutions to exchange the ability of identifying wider optima. In contrast, our method further improves the capability of identifying wider minima over DML whilst maintaining the original exploring quality.

Variance Analysis on ONE's Branches

We analysed the variance of ONE's branches over the training epochs in comparison to the conventional ensemble method. We used ResNet-32 as the base net and tested CIFAR100. We quantified the model variance by the average prediction differences on training samples between every two models/branches in Euclidean space. Figure 5.8 shows that a 3-Net Ensemble involves *larger* inter-model variances than ONE with 3 branches throughout the training process. This means that the branches of ONE have higher correlations, due to the proposed learning constraint from the distillation loss that enforces them align to the same teacher prediction, which probably hurts the ensemble performance. However, in the mean generalisation capability (another fundamental aspect in ensemble learning), ONE's branches (the average error rate 26.61±0.06%) are much superior to individual models of a conventional ensemble (31.07±0.41%), leading to a stronger ensembling performance.

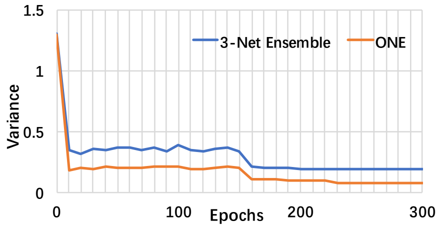


Figure 5.8: Model variance during training.

5.3 Summary

This chapter focuses on the cross-model knowledge transfer on image classification. To solve the expensive cost of training teacher model and multi-stage process for knowledge distillation, two methods are proposed. First, a novel Self-Referenced Deep Learning (SRDL) strategy is proposed for improving deep network model learning by exploiting self-discovered knowledge in a two-stage training procedure. SRDL can train more discriminative small and large networks with little extra computational cost. This differs from conventional knowledge distillation which requires a separate pre-trained large teacher model with huge extra computational and model training time cost. Conceptually, SRDL is a principled combination of vanilla model optimisation and existing knowledge distillation, with an attractive trade-off between model generalisation and model training complexity. Extensive experiments show that a variety of standard deep networks can all benefit from SRDL on both coarse-grained object categorisation tasks (image classification) and fine-grained person instance identification tasks (person re-identification). Significantly, smaller networks benefit from more performance gains, making SRDL specially good for low-memory and fast execution applications. The further component analysis gives insights to the SRDL's model design considerations. Second, a novel On-the-fly Native Ensemble (ONE) strategy is proposed for improving deep network learning through online knowledge distillation in a one-stage training procedure. With ONE, we can more discriminatively learn both small and large networks with less computational cost, beyond the conventional offline alternatives that are typically formulated to learn better small models alone. Our method is also superior over existing online counterparts due to the unique capability of constructing a high-capacity online teacher to more effectively mine knowledge from the training data and supervise the target network concurrently. Extensive experiments on four image classification benchmarks show that a variety of deep networks can all benefit from the ONE approach. Significantly, smaller networks obtain more performance gains, making our method specially good for low-memory and fast execution scenarios.

Chapter 6

Knowledge Transfer Across Tasks in Image Classification

Often, we are too slow to recognize how much and in what ways we can assist each other through sharing expertise and knowledge.

—— **Owen Arthur**

This chapter discusses the knowledge transfer across tasks in image classification. Specifically, few shot classification tasks are investigated. The knowledge transfer process across the task is through the prototype, which is the global representation of the category. To address the drawbacks associated with the common practice of employing a single prototype per class, this chapter proposes a mixture prototype framework for metric based approach. Furthermore, this chapter also considers the influence of unlabelld samples to the cross-task knowledge transfer by developing a novel fewmatch methods.

6.1 Learning Diverse and Representative Prototype Mixtures for Few-shot Classification

6.1.1 Methodology

Let us consider a training dataset D_{base} with annotated samples $X_b = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ and their corresponding labels $Y_b = \{y_1, ..., y_n\}$, comprising C_b base categories. Our test dataset D_{novel} contains C_n novel classes, each of which is associated with only a few labelled samples (e.g. ≤ 5 samples), while the remaining unlabelled samples are used for evaluation. The goal of few-shot classification is to learn a classifier on D_{base} that can generalise well to the C_n novel classes based on the limited labelled samples from C_n novel categories. Specifically, these labelled samples constitute the *support set* S_n with K_n annotated samples per class, while the unlabelled samples form the query set Q_n on which the model is evaluated. This is also referred to as a C_n-way K_n-shot classification problem. A large set of FSL methods also use the concept of episode training, sampling subsets of *support* S_b and *query* Q_b sets from D_{base} in order to mimic the support-query test scenario.

We propose a generic mixture of prototypes strategy that can be integrated with metric-learning based methods. An overview of the proposed method is provided in Figure 6.1 where we provide integration of our method with imprinted weights FSL method [69] as an example. Our approach focuses on learning high quality prototypes and maximally leveraging the use of multiple class-specific representations. We augment a global image feature representation with a set of N local representations focusing on distinct regions through the use of local and global average pooling. These representations, computed on the support set, constitute the class prototypes that are subsequently used to classify unlabelled examples using the cosine distance. This enables the exploitation of high-granularity local descriptors without sacrificing global information. We find that prototypes, obtained from local image input, may be of poor quality if they focus on ambiguous or irrelevant image regions (e.g. background). We address this issue using a self-supervised rotation loss to learn robust features, and a soft attention gate to combine prototype classification decisions.

Mixture of prototypes formulation

Metric-based FSL methods focus on learning strong feature representations θ_f, which regroup images of the same class and separates different classes with respect to a pre-defined distance metric $\gamma(\cdot)$. Depending on the method considered, a prototype p_c associated with class c, can be defined during training as either (a) the average representation of support set images S_c (episodic training methods [62]) or (b) the c^{th} column of classifier weights trained via standard backpropagation on the base dataset [69]. At test time, all methods employ option (a). Unlabelled images x are then classified based on their embedding distance to the different class prototypes $\gamma(x, p_c)$. A large body of work has focused on developing better training strategies and architectures, including nearest-neighbour based episode training [62, 61, 63] and training a cosine classifier on the whole base dataset [69]. Nonetheless, very few works have attempted to alleviate the inherent

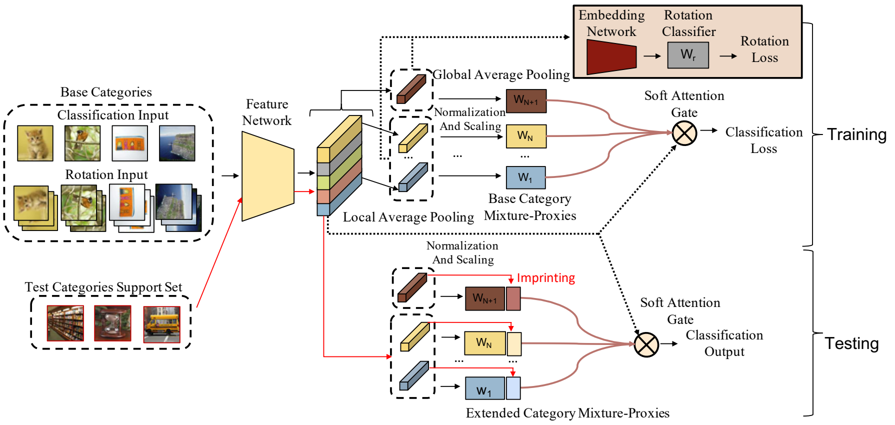


Figure 6.1: An overview of the proposed MP model with the imprinted weights implementation. Given a set of training samples from the base categories, we learn a set of diverse feature representations via global and local pooling, each of which is associated with a trainable classifier. Classification decisions are made based on the scaled cosine distance between the normalized input embeddings and the columns of the classifier weight matrices W_i such that each column of W_i constitutes a trainable class prototype. Local prototype representations are optimised using a self-supervised rotation loss associated with a rotation specific embedding network, and a soft attention gate selects the highest quality prototypes to obtain an ensemble classification decision. At test time, prototypes for new classes are computed by averaging representations over the support set and imprinted in the trained classifiers, effectively allowing to test new classes without retraining.

bias and limitations linked to the use of a single representation.

Our objective is to learn a richer category representation using a mixture of prototypes to accurately represent the variability within one class. We propose to decompose the support set representation into a set of $N+1$ prototype representations $\{p_c^n\}$, $n \in [1, \ldots, N+1]$ each of which will make individual distance based class assignments. We carefully design our model so as to maximally leverage multiple prototypes through the use of both local and global model component considerations: enforcing high variance, employing an auxiliary task using image rotation to increase robustness to local inputs and improve local spatial reasoning and a soft attention gate to increase the influence of reliable prototype predictions.

Learning highly diverse prototypes

An important criterion for the design of our mixture of prototypes is to maximise the ***variance*** between prototypes so as to minimise redundancy between the representations. To this end, we propose a local and global prototype learning method. Considering an annotated image $\boldsymbol{x}^b$ from the training dataset D_{base}, we denote $\theta_f(\boldsymbol{x}^b)$ as its representation, where $\theta_f(x^b) \in \hat{F} \times \hat{W} \times \hat{H}$, and $\hat{F}$, $\hat{W}$, $\hat{H}$ are the feature vector channel, width and height respectively. Instead of simply using the whole image for average pooling, we firstly use average pooling on N disjoint local regions which are obtained by uniformly partitioning the image feature representation along its height H, width W or both such that the n^{th} local prototype will focus on a specific region R_n of the input image. The number of prototypes, defined by height and / or width partitioning, constitutes a hyperparameter.

By designing local prototypes that focus on disjoint parts of the image, we force prototypes to provide complementary information and limit redundancy. However relying solely on fine-grained, local representations would disregard global, high level information that can also provide highly useful cues. As a result, we combine our set of local prototype representations $p_n, n \in [1, \ldots, N]$ with a global prototype p_{N+1} that considers the whole image, computed in parallel by global average pooling of $\theta_f(\boldsymbol{x}^b)$. This combination of local and global descriptors enables computation of a set of diverse class prototypes that focus on different aspects of the image. However, a naïve use of multiple local descriptors results in two important problems that can limit the performance of multi-prototype strategies. Firstly, learning accurate embeddings and classifiers using local prototypes can be challenging and obtains subpar performance, due to the potential ambiguity associated with partial image inputs. Secondly, local prototypes may focus on non-discriminative image regions and therefore provide no relevant information, hurting overall performance. We address these shortcomings as following section.

Regularising local prototypes with self-supervision

Recent advances in unsupervised and semi-supervised learning have demonstrated the advantage of self-supervision to regularise model training and learn stronger feature representations [176]. Training classifiers using local image information provides a scenario with an analogous challenge, where local information can be ambiguous or may not even contain the class of interest. This potentially unreliable signal hurts model training and may yield sub-optimal prototype representations. Integration of a self-supervised auxiliary task allows the learning of more robust features, and therefore prototypes, by extracting features suitable for multiple high level tasks. This effectively allows to optimise the local prototypes' ***representative*** power. Self-supervision has recently been applied successfully in the single prototype few-shot learning scenario [253] in the form of rotation and jigsaw puzzle tasks.

In our scenario, we consider an auxiliary rotation task suitable, as rigid rotation retains spatial contiguity and image properties helpful to our main task, unlike other common alternatives *e.g.* jigsaw puzzle tasks. Formally, given an image x^b from D_b, we produce four rigidly transformed images by rotating x^b by r degrees, where $r \in \{0°, 90°, 180°, 270°\}$. We formulate the auxiliary rotation task as a four class classification problem, where the objective is to correctly recognize rotation r. This is achieved by training a linear classifier W_r after passing image local embeddings $\theta_f(x_i^b)_n, n \in [1, \ldots, N]$ and global embedding $\theta_f(x_i^b)_{N+1}$ through a 1×1 convolution layer. This additional convolutional layer adapts the feature vector $\theta_f(x_i^b)$ to the rotation task and additionally implicitly discourages conflict with the main classification task. The rotation branch is finally trained using a standard softmax cross-entropy loss:

$$\mathcal{L}_{\text{rotate}} = -\frac{\sum_{i=1}^{\text{N+1}} \sum_{c=1}^{4} \delta_{c,y} \log \Big(\rho_c \big(\Phi(\theta_f(\boldsymbol{x})_i) \big) \Big)}{\text{N+1}}, \tag{6.1}$$

where Φ is the rotation embedding function, ρ_c is the rotation prediction score and $\delta_{c,y}$ is the Dirac delta function.

Ensembling prototype predictions with attention

Recall that local prototype classification task utility will vary; we propose to learn this and weight prototype ensembles using attention. For a given input image x, prototype-specific classification scores $f_n(x)$ are associated to image region R_n, and are computed as the normalised distance between the embedding of $\theta_f(x)_n$ and prototypes p_n of all C_N classes:

$$f_n^c(x) = \frac{\exp\Big(\big(p_n^c, \theta_f(\boldsymbol{x})_n \big) \Big)}{\sum_{j=1}^{C_N} \exp\Big(\big(p_n^j, \theta_f(\boldsymbol{x})_n \big) \Big)} \tag{6.2}$$

where f_n^c and p_n^c are, respectively, the classification score and prototype associated with class c.

A straightforward strategy would involve averaging all prototype decisions to obtain an ensemble global score. However, such a strategy is at risk of being significantly affected by uninformative local prototypes focusing on non-discriminative regions, as discussed in Sec. 6.1.1. We alternatively choose to integrate a soft attention gate, thus modulating the combination of prototype decisions and affording attenuation of the signal propagated by low quality prototypes.

We design the soft attention gate $\mathcal{G}$ as a single softmax and fully connected layer, taking as input the global image representation $\theta_f(x)$, reshaped into a vector. The attention weight of each prototype $\alpha{=}\{\alpha_n\}$ can then be calculated as $\alpha = \mathcal{G}(\theta_f(x)){+}1$. To mitigate potential errors induced by noisy or difficult examples, we follow [254] and combine our gate with a residual connection, yielding more robust performance to inaccurate attention weights. Finally, classification scores for image x are computed as:

$$f(x) = \frac{\sum_{n=1}^{N+1} \alpha_n f_n(x)}{N+1} \tag{6.3}$$

The model's classification branch can then be trained using the predictions and standard metric learning strategies.

Mixture of prototypes with imprinted weights

Our mixture of prototypes model provides a general formulation that can easily be integrated in conjunction with popular metric-based few-shot learning models. To exemplify, we investigate implementation with the imprinted weights model of [69]. We note that other popular episode training strategies [61, 62] also constitute valid options. The imprinted weights approach trains a classifier on the whole set of base classes C_b. The architecture comprises a feature extraction network θ_f, followed by a classifier consisting of fully connected layer without bias $W \in F \times C_b$, where F is the output dimension of θ_f. The key idea is to learn W such that the cosine distance between w_c (the c^{th} column of W) and the embedding $\theta_f(x_c)$ of input images of class c is minimal. Thus, w_c can be seen as the *prototype* of the c^{th} category in the base set. The objective function aims to minimise the cosine distance between images and their corresponding prototype.

The imprinted weights model provides two main advantages. Firstly, due to the training strategy, each row of the classifier matrix W constitutes a prototype, allowing new categories to easily be imprinted in W using the support set prototype. This alleviates the need to retrain or fine-tune a model when new categories are available or when the number of shots is changed, yielding a highly efficient model with continual learning ability. Second, the classifier training approach does not require a cumbersome episodic training process. However, the imprinting strategy makes the model highly sensitive to prototype quality and easily fails in the single prototype scenario. Our approach focuses on strong multi-modal representations and allows full exploitation of the benefits of this model while maintaining robust performance. In this context, we can integrate our mixture of prototypes approach in a very natural way, associating each of the N local and single global feature vectors with a different classifier. Classification decisions are computed by evaluating the cosine distance between an input image and each column of a given classifier matrix, where a column corresponds to a class. As such, classifier weights are learned to minimise the distance between embeddings and prototypes (classifier columns) of the same class. As each classifier focuses on different feature regions, we are able to automatically learn our $N+1$ diverse local prototypes (and global prototype) as columns of each classifier matrix, $W_1, W_2, ..., W_{N+1}$. Specifically, for a given classifier W_i, the classification score of sample x for class c is computed as:

$$f_i^c(x) = \frac{\exp\left(\gamma\left(w_{ic}^T, \theta_f(\boldsymbol{x})_i\right)\right)}{\sum_{j=1}^{C_b} \exp\left(\gamma\left(w_{ij}^T, \theta_f(\boldsymbol{x})_i\right)\right)} \tag{6.4}$$

where w_{ij} is the j^{th} column of weight matrix W_i and corresponds to prototype p_{ij} associated with region R_i and class j. The scaled cosine similarity is defined as $\gamma\left(w_j^T, \theta_f(\boldsymbol{x})\right) = s w_i^T \left(\theta_f(\boldsymbol{x})\right)$. Both W_i and $\theta_f(\boldsymbol{x})$ are normalized using the L_2 norm, and s is a trainable scalar, introduced in [69] to avoid the risk that the cosine distance yields distributions that lack discriminative power. Then, the classification loss function is calculated as follows:

$$\mathcal{L}_{\text{ce}} = -\frac{\sum_{c=1}^{C_\text{b}} \delta_{c,y} \log f^c(x) + \sum_{n=1}^{N+1} \sum_{c=1}^{C_\text{b}} \delta_{c,y} \log f_n^c(x)}{N+2} \tag{6.5}$$

where f^c is computed from all f_i^c using Eq (6.3) and $\delta_{c,y}$ is the Dirac delta function. We purposefully retain a summation of individual $\log f_i^c(x)$ terms in Eq (6.5) to ensure that each prototype is pushed to possess discriminative class information. The whole model can then be trained end-to-end using the objective function $\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{rotate}}$.

At test time, given a new category j from D_{novel} with support dataset $S = \{\boldsymbol{x}_1^s, y_1^s, \ldots, \boldsymbol{x}_n^s, y_n^s\}$, we can compute a new set of prototypes as:

$$p_{nj}^* = \frac{1}{|S_j|} \sum_{\boldsymbol{x}_i^s \in S_j} \theta_f(\boldsymbol{x}_i^s)_n, \forall n \in [1, \ldots, N+1] \tag{6.6}$$

where S_j contains all annotated samples in the j^{th} category. By imprinting classifier W_n^* with $w_{nj}^* = p_{nj}^*$ and repeating the process for any new category, we are able to recognise new classes without retraining the model. By concatenating W_n and W_n^*, we are able to test on all $C_n + C_b$ categories.

Table 6.1: The mean accuracies of the 5-way 1-shot and 5-shot tasks on the *mini*ImageNet and *tiered*ImageNet. Multi-prototype methods are highlighted in gray. M: Metric, G: Gradient. †: Results on *tiered*ImageNet as reported in [255] (**Red/Blue: Best and second best results on 4Conv or ResNet/WRN-28-10 backbones.**)

Model	Backbone	Type	*mini*Imagenet		*tiered*Imagenet	
			1-shot	5-shot	1-shot	5-shot
Baseline++[256]	*4Conv*	M	48.24 ±0.75	66.43 ±0.63	-	-
MatchingNet [61]	*4Conv*	M	43.56 ±0.84	55.31 ±0.73	-	-
ProtoNet† [62]	*4Conv*	M	49.42 ±0.78	68.20 ±0.66	53.31 ±0.89	**72.69 ±0.74**
RelationNet † [63]	*4Conv*	M	50.44 ±0.82	65.32 ±0.70	**54.48 ±0.95**	71.32 ±0.78
IMP [74]	*4Conv*	M	49.20 ±0.70	64.70 ±0.70	-	-
DN4 [75]	*4Conv*	M	51.24 ±0.74	71.02 ±0.64	-	-
DYNAMIC-FSL† [175]	*4Conv*	M	**56.20 ±0.86**	**73.00 ±0.64**	50.90 ±0.46	66.69±0.36
MAML [65]†	*4Conv*	G	48.07 ±1.75	63.15 ±0.91	51.67 ±1.81	70.30 ±1.75
iMAML HF [67]	*4Conv*	G	49.30 ±1.88	-	-	-
LCC [167]	*4Conv*	G	54.60 ±0.40	71.10 ±0.40	-	-
Ours	*4Conv*	M	**57.13 ±0.78**	**74.25 ±0.62**	**62.96 ±0.91**	**79.13 ±0.63**
TADAM [257]	*ResNet-12*	M	58.50 ±0.30	76.70 ±0.30	-	-
DC [258]	*ResNet-12*	M	62.53 ±0.19	78.95 ±0.13	-	-
TapNet [259]	*ResNet-12*	M	61.65 ±0.15	76.36 ±0.10	63.08 ±0.15	80.26±0.12
ECMSFMT[260]	*ResNet-12*	M	59.00	77.46	63.99	81.97
CTM [255]	*ResNet-18*	M	62.05 ±0.55	78.63 ±0.06	64.78 ±0.11	81.05 ±0.52
wDAE-GNN [261]	*WRN-28-10*	M	**62.96 ±0.15**	78.85 ±0.10	-	-
PPA [64]	*WRN-28-10*	M	59.60 ±0.41	73.74 ±0.19	-	-
CCrot[179]	*WRN-28-10*	M	62.93 ±0.45	**79.87 ±0.33**	**70.53 ±0.51**	**84.98 ±0.36**
CAML [262]	*ResNet-12*	G	59.23 ±0.99	72.35 ±0.18	-	-
MTL [168]	*ResNet-12*	G	61.20 ±1.80	75.50 ±0.80	-	-
MetaOptNet-SVM [169]	*ResNet-12*	G	62.64 ±0.61	78.63 ±0.46	65.99 ±0.72	81.56 ±0.53
LEO [68]	*WRN-28-10*	G	61.76 ±0.08	77.59 ±0.12	66.33 ±0.05	81.44 ±0.09
Ours	*ResNet-12*	M	**66.17±0.75**	**82.40±0.57**	**71.95±0.92**	**86.74±0.61**

6.1.2 Experiments

Experimental Set-up

We evaluated our model on two popular FSL benchmarks: ***mini*ImageNet** and ***tiered*ImageNet**. *mini*ImageNet [61] is a subset of the ImageNet ILSCVRC-12 [3] dataset, consisting of 60000 images uniformly distributed over 100 classes. We use the standard 64/16/20 classes split for train / val / test as proposed in [263]. *tiered*ImageNet [183] is another subset of ILSVRC-2012, specifically designed to increase the semantic dissimilarity between the different category splits. It contains 608 classes consisting of 34 high-level categories. These are divided into 20/6/8 coarse categories for train / val / test splits with 351, 97 and 160 classes, respectively. To generate validation and test episodes, we follow the strategy adopted in [175]. We randomly sample C_n classes from the validation or test set and we sample K_n labelled images from each class as support images, and 15 images as query images. We report the mean accuracy of 600 randomly generated test episodes with 95% confidence intervals.

The method was implemented using PyTorch [264]. We use ResNet-12 [257] as the backbone architecture for the embedding network due to its strong performance. We use the training protocol and parameters described in [169, 175]: our model is optimised using SGD with momentum set to 0.9, weight decay to 0.0005, mini-batch size to 256, and 60 epochs. All input images were resized to 84×84. The learning rate was initialised to 0.1, and updated to 0.006, 0.0012, and 0.00024 at epochs 20, 40 and 50, respectively. Following [169], we use DropBlock regularization [265]. We initialized the trainable scalar s to 10, and used five local prototypes along the height dimension and one global prototype. Our baseline in subsequent experiments constitutes the single prototype imprinted weights model with identical common training parameters. Ablation experiments are carried out on the 5-way 1 shot scenario on miniImageNet unless specified otherwise.

Comparison to State-of-the-Art Methods

Table 6.1 shows a comparison between our method and state of the art approaches, including **(a)** 15 metric based models and **(b)** 7 meta-gradient based approaches. To decouple the influence of the method from the embedding backbone on the obtained results, we indicate which backbone network is used by each method and report our performance using ResNet-12 and 4Conv backbones. Based on Table 6.1, we observe the following: **(1)** metric learning and meta-gradient based approaches have similar performance, with wDAE-GNN, CCrot (metric

Table 6.2: Classification accuracy on *mini*ImageNet of standard FSL methods with and without our MP approach.

MODEL	METRIC(%)	
	1-SHOT	5-SHOT
PROTONET	59.25 ±0.87	75.13 ±0.65
PROTONET+**MP**	**62.52 ±0.86**	**76.22 ±0.69**
MATCHINGNET	59.55 ±0.86	72.34 ±0.65
MATCHINGNET+**MP**	**62.10 ±0.87**	**75.86 ±0.66**

based) and MetaOptNet-SVM (meta-gradient) achieving similar, state of the art performance on *mini*ImageNet and *tiered*ImageNet. **(2)** Models' backbones vary substantially across methods, yielding large gaps in performance, in particular between 4Conv and other, more complex backbones. **(3)** we observe that our model significantly outperforms the state of the art by a large margin. Specifically, we significantly outperform multi-prototype approaches IMP and DN4; and surpass the strongest approaches wDAE-GNN, MetaOptNet-SVM and CCrot in terms of accuracy by over 3% (*mini*ImageNet) and 1.5% (*tiered*ImageNet) in both 1-shot and 5-shot settings.

Ablation Experiments

Integration in other metric-learning methods. Our mixture of prototypes is a generic method that can be easily integrated with other metric learning based approaches. To evaluate the versatility of our approach, we explore integration with both ProtoNet [62] and MatchingNet [61] using default parameters. These approaches are highly popular and constitute the backbone of a large set of metric-based FSL methods. Table 6.2 reports the classification accuracy for both methods, with and without integration of our MP. We see that the performance of both methods improve when using our approach, with an improvement of up to 3%, highlighting the modularity and advantage of our formulation.

Model design analysis. We evaluate the influence of each component of our model on overall performance to understand and attribute component credit. Namely we evaluate the local / global Mixture of Prototypes (MP), the Rotation Auxiliary task (Rot), the soft attention gate (Gate), and the rotation Embedding Network (EN). Table 6.3 shows that each component makes a clear contribution to the performance gain. With the full model, we achieve significant performance improvement over the baseline: 6.43%, 4.99% for 1-shot on *mini*ImageNet and *tiered*ImageNet repectively.

Table 6.3: Model design analysis on *mini*Imagenet. MP: Local/global mixture of prototypes, Rot: auxiliary rotation task, Gate: soft attention gate, EN: rotation embedding network.

Model Components				*mini*ImageNet		*tiered*ImageNet	
MP	Rot	Gate	EN	1-shot	5-shot	1-shot	5-shot
				59.74 ±0.82	76.65 ±0.60	66.96 ±0.92	83.18 ±0.65
✓				62.81 ±0.81	79.43 ±0.61	69.66 ±0.94	85.05 ±0.64
	✓			62.73 ±0.78	78.67 ±0.58	70.00 ±0.87	85.73 ±0.60
✓	✓		✓	65.79 ±0.78	81.17 ±0.56	71.67 ±0.83	86.52 ±0.58
✓	✓	✓		61.56 ±0.77	79.57 ±0.60	69.42 ±0.91	85.85 ±0.58
✓	✓	✓	✓	**66.17 ±0.75**	**82.40 ±0.57**	**71.95 ±0.92**	**86.74 ±0.61**

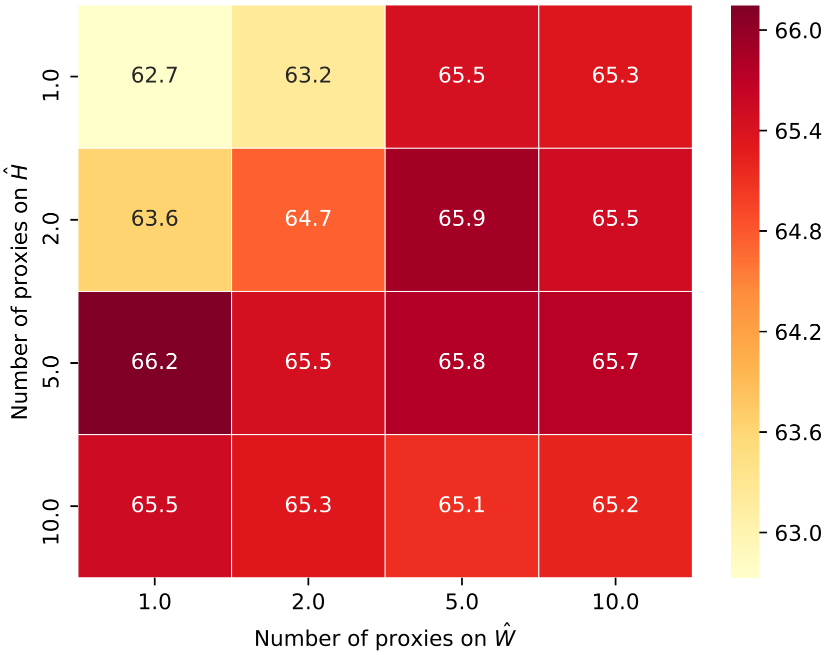


Figure 6.2: Influence of the number of prototypes along the image's feature width ($\hat{W}$) and height ($\hat{H}$) on classification accuracy.

The design strategy of prototypes. We further evaluated the influence of two components of our prototype design strategy: a) the number of prototypes and the regions on which they focus and b) the influence of the rotation task on local prototype representative power. Regarding a) we train models with a varying number of prototypes along the height and width of the image feature.

Results are shown in Figure 6.2. We consider 1, 2, 5 or 10 prototypes along each dimension. We find that performance is not influenced by the direction (height or width) on which prototypes are computed, but rather by the overall number of prototypes. We verify that prototype count can be tied to the expressive power of the model and note that accuracy appears to be stable when the total number of prototypes is $>= 5$, exhibiting model stability with respect to the design strategy. Regarding b) we evaluate the influence of the rotation loss on the classification performance of each individual prototype with and without use of the auxiliary task for two separate sets of 600 test episodes. We report the performance gain in terms of accuracy in Figure 6.3 and show that, as hypothesised, our self-supervision task mainly improved local prototype performance (+4% on average) in comparison to global prototypes (+2%). This is observed consistently across both sets of episodes.

Free shot setting Since our imprinted weights based MP implementation does not require episode training, our trained model is independent of the number of shots at test time. We compared the single prototype baseline to our MP method in the free shot setting in Figure 6.4 for both ResNet and 4Conv backbones. We observe that both baseline and our MP model consistently improve with the number of shots increase, showing that our formulation is robust to the number of shots and that our model surpasses the baseline for all configurations.

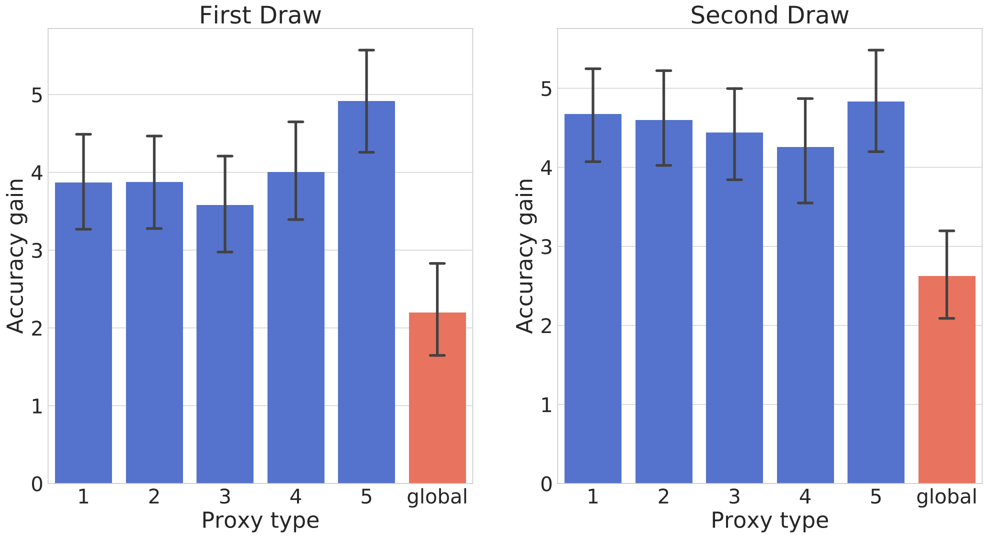


Figure 6.3: Accuracy gain on local (blue) and global (red) prototypes using the rotation task. Two 600 random test episode draws.

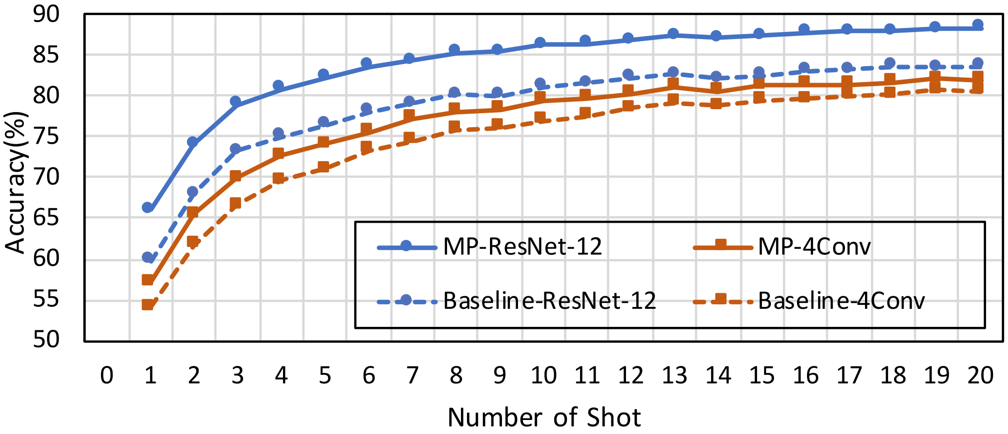


Figure 6.4: Free shot evaluation on *mini*Imagenet.

6.2 Fewmatch: Dynamic Consistency Regularization for Semi-Supervised Learning

In reality, we often have access to a large amount of unlabelled data without expensive human labelling cost. However, the proposed approach in the above section ignores the latent information from unlabelled data. In this section, we investigate the benefit of utilizing unlabelled data on cross-task knowledge transfer.

6.2.1 Methodology

Similar to Section 6.1, we consider a base training dataset $D_{base} = \{X_b^l, X_b^u\}$ comprising Labelled Data (LD) $X_b^l = \{\boldsymbol{x}_1^l, \dots, \boldsymbol{x}_n^l\}$ with labels $Y_b = \{y_1, \dots, y_n\}$, while the difference is an additional set of Unlabelled Data (UD) $X_b^u = \{\boldsymbol{x}_1^u, \dots, \boldsymbol{x}_m^u\}$ is provided in this section. All examples in D_{base} belong to one of C_b base categories. Our novel dataset D_{novel} contains C_n disjoint novel classes each with only a handful of labelled samples (*e.g.* ≤ 5) as well as a further limited set of unlabelled samples (*e.g.* ≤ 100) per class with which to fine-tune the model. D_{novel} further comprises unlabelled samples used for evaluation. Our objective, similarly to standard few-shot settings, is to learn a classifier capable of accurately recognising novel classes, despite having only a limited amount of available LD. However in contrast to standard FSL, we possess additional UD for both base and novel classes, which we aim to leverage in order to maximise performance. To formalise our setting, we consider that D_{novel} comprises of a fixed *support set* of K_n^l labelled and K_n^u unlabelled examples per class, and refer to the remaining unlabelled test images as the query

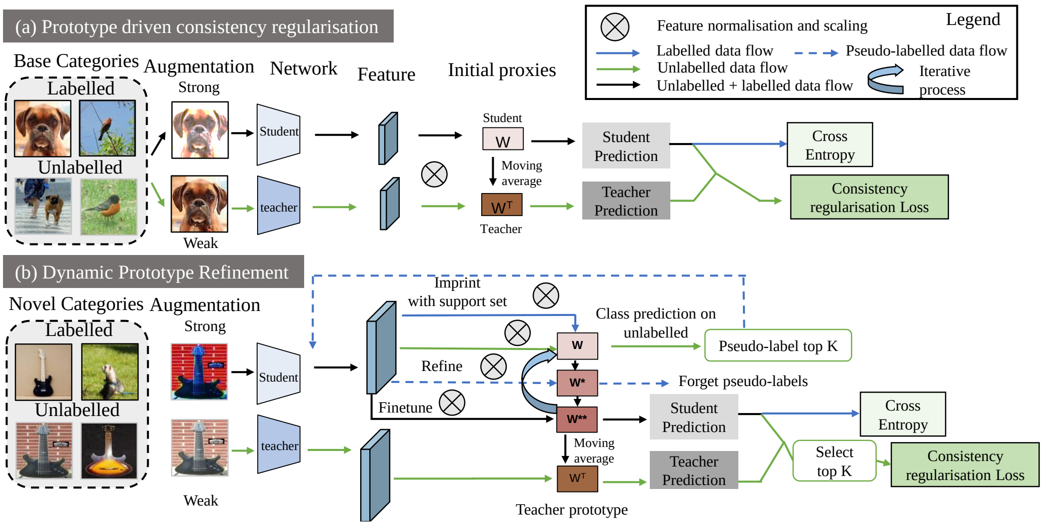


Figure 6.5: Overview of the proposed FewMatch approach. See main body for details.

set Q_n. This C_n-way K_n^l-shot classification problem defines a standard SS-FSL setting.

An overview of our proposed “FewMatch” method is provided in Figure 6.5. FewMatch first trains a classification model on D_{base} by exploiting the concept of imprinted weights (IW). IW allow end-to-end model training, while at the same time learning global class feature representations (commonly referred to as prototypes [62]) utilised as classifier weights. This is achieved by computing predictions as the cosine similarity between input features and classifier. End-to-end training allows seamless introduction of a CR loss, effectively leveraging UD to train a strong feature extractor and learn high quality prototypes. The second stage involves model fine-tuning on D_{novel} in order to leverage UD for novel classes. Our cosine classifier formulation, which minimises distance between input feature and classifier weights, enables the initialisation of classifier weights for new, previously unseen classes with the average feature vector of the novel labelled support set. This *strong* initialisation enables exploitation of UD and allows our CR approach to dynamically refine prototypes via a two-level scheme. Prototypes are updated iteratively using a combination of explicit feature averaging and implicit parameter updates. This is achieved by 1) extending the support set using unlabelled samples with highest prediction confidence and 2) fine-tuning the model using a CR scheme.

Prototypes Driven Consistency Regularisation

Our formulation exploits the popular concept of prototypes as discussed in Section 6.1 , Prototypes $\mathcal{P} = \{p_1, p_2, \dots, p_{C_b}\}$ are learned global feature representations, each describing a particular class to recognise. Class prototypes are typically defined as the average feature representation of the support set. They are learned such that the distances between input samples, of a given class, and the respective class prototype is minimised (else maximised). Our prototype-based formulation exploits the concept of imprinted weights [69], allowing prototypes to be learned using a classifier trained end-to-end on the entire set of base classes C_b. In this section we firstly introduce our imprinted weight formulation and then describe the integration of this within a teacher-student framework, enabling the introduction of our CR loss.

Imprinted weights formulation. Our classification model uses a standard architecture, comprising a feature extraction network θ_f, and a classifier defined by a fully connected layer without bias $W \in F \times C_b$, where F is the output dimension of θ_f. The main idea of imprinted weights is to train the model such that, for a given class c, the cosine similarity between the embedding vector $\theta_f(\boldsymbol{x})$ of input image $\mathbf{x}$ and the corresponding column w_c of W is maximised. By normalising the classifier and embedding vectors, the model can be trained end-to-end using a standard cross entropy loss. In this setting, w_c is regarded as the *prototype* representation of class c and can be learned implicitly without the typically required, episode training strategy and support set averaging. More formally, for input sample $\mathbf{x}$, the set of classification scores output by the model is $f(x) = \{f^1(x), f^2(x), \ldots, f^c(x), \ldots, f^{C_b}(x)\}$ and the score $f^c\big(\mathbf{x}\big)$ for a given class c can be computed by

$$f^c\big(\mathbf{x}\big) = \frac{\exp\Big(\gamma\big(\mathbf{w_c^T}, \theta_\mathbf{f}(\boldsymbol{x})\big)\Big)}{\sum_{\mathbf{i=1}}^{\mathbf{C_b}} \exp\Big(\gamma\big(\mathbf{w_i^T}, \theta_\mathbf{f}(\boldsymbol{x})\big)\Big)} \tag{6.7}$$

where $\mathrm{w_i}$ is the i^{th} column of weight matrix W and the prototype p_i of class i. The scaled cosine similarity is then given by $\gamma\big(\mathrm{w_i^T}, \theta_\mathrm{f}(\boldsymbol{x})\big) = s \cdot \mathrm{w_i^T}\big(\theta_\mathrm{f}(\boldsymbol{x})\big)$. $\mathrm{w_i}$ and $\theta_\mathrm{f}(\boldsymbol{x})$ are normalized using the L_2 norm, and s is a trainable scalar, as introduced by [69] to avoid the risk that the cosine distance yields distributions lacking in discriminative power.

Finally, the classification loss can be calculated as: $\mathcal{L}_{\mathrm{ce}}(\mathbf{x}) = -\sum_{\mathbf{c=1}}^{\mathrm{C_b}} \delta_{\mathbf{c,y}} \log \mathbf{f^c}(\mathbf{x})$ where $\delta_{c,y}$ is the Dirac delta function. Defining class prototypes as learnable model weights affords end-to-end training and enables introduction of CR to our model in a natural fashion. These decisions allow us to leverage UD and implicitly refine prototypes without explicit pseudo-labelling. Furthermore, this approach optimises the base class learning process by allowing full exploitation of the available LD without the typical requirement that necessitates simulation of the few-shot set-up (episode training) [65].

Consistency Regularisation. We highlight that the described training strategy does not yet leverage UD, available in the considered SS-FSL problem setting. Towards taking advantage of UD, we introduce a CR loss [76] that is driven by the learned prototypes. The idea underlying CR is to regularise predictions such that they become invariant to small input perturbations that do not affect class semantics. This strategy has been used successfully for a variety of problems and is particularly appealing in the semi-supervised context as it leverages UD without explicit pseudo-labelling. A key difference in our setting, with respect to conventional SSL, is that our CR loss directly depends on prototype instantiations, as predictions are based on the distance between input and each class prototype. This strategy drives our approach to learn more discriminative and robust prototypes so as to maintain classification accuracy under different input perturbations.

Following strategies adopted in the recent SSL state of the art [76, 79], we embed our IW model within a teacher-student framework [76] where we seek to impose consistency between teacher and student predictions. Both teacher and student networks share the same architecture, however only student weights are optimised by back-propagation. Teacher weights θ^T are alternatively computed as an Exponential Moving Average (EMA) of the student weights θ , $\theta^T = (1-\alpha)\theta^T + \alpha\theta$. Such temporal averaging strategies have been shown to yield more robust and accurate models and are therefore desirable in the often noisy few-shot setting.

Considering an unlabelled sample u_b we realise sample perturbations, as suggested in [266, 79], by generating $\bar{u}_b$ and $\hat{u}_b$ using *weak* and *strong* augmentations respectively. The weak augmentation sample $\bar{u}_b$ has the goal of improving prediction stability in the teacher network. This strategy helps to constrain the strong augmentation sample prediction. The consistency loss is then computed as:

$$\mathcal{L}_{\text{cons}}(u_b) = ||\text{Sharp}(f_t(\hat{u}_b), \mathcal{T}) - f_s(\bar{u}_b)||^2, \qquad \text{where} \quad \text{Sharp}(x, \mathcal{T}) = \frac{x_i^{\frac{1}{\mathcal{T}}}}{\sum_{j=1}^{C_b} x_j^{\frac{1}{\mathcal{T}}}} \tag{6.8}$$

such that f_s and f_t are predictions computed by the student and teacher networks respectively; and Sharp$(\cdot)$ is a sharpening function introduced as in [78] to reduce the entropy of the label distribution. In summary, the model is trained on the base classes using the global loss function $\mathcal{L}_{base} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{cons}}$, where the hyperparameter λ balances the relative influence of the two terms.

Dynamic Prototype Refinement

Our training stage, using D_{base}, yields a model capable of estimating reliable class prototypes on novel, unseen categories. In a standard few-shot setting (*i.e.* without available UD), prototypes are often estimated directly from the support set and reliable performance can be achieved without further training. In our problem setting, we set the objective of exploiting the additionally available UD in order to obtain strong prototype initialisations that can then lend themselves to further refinement. Towards this goal, the second component of FewMatch constitutes our Dynamic Prototype Refinement (DPR) strategy to take advantage of the UD, available from D_{novel}, towards improving model adaption to novel categories. The approach we introduce is, by design, able to improve performance on novel categories despite the presence of limited data regimes. DPR comprises three stages: **(1)** Prototype Initial Inference (PII), via the introduced IW procedure **(2)** *Explicit* prototype refinement using top-K selection and **(3)** *Implicit* prototype refinement using CR. Prototypes are initially estimated during the first step and then dynamically updated iteratively using steps two and three, such that prototype quality is improved. The remainder of this section provides further detail on steps **(1)**-**(3)** and the iterative process. Algorithm 3 details pseudocode for the complete DPR process.

(1) Prototype Initial Inference. Given new category j from D_{novel} with support set S_j=$\{\boldsymbol{x}_1^s, y_1^s, \ldots, \boldsymbol{x}_n^s, y_n^s\} \cup \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, compute an initial prototype as:

$$p_j^* = P(S_j) = \frac{1}{|S_j|} \sum_{\boldsymbol{x}_i^s \in S_j} \theta_f(\boldsymbol{x}_i^s), \tag{6.9}$$

The estimated prototype is then imprinted in classifier W as $w_j = p_j$ and the process is repeated for each new category (see Figure 6.5). This allows for recognition of new classes without model retraining and provides high quality initialisations for our dynamic refinement stage.

(2) Explicit Prototype Refinement. We highlight that initial prototypes, computed using Eq. 6.9, do not make use of the additional UD available for novel classes. Exploiting UD can be considered crucial for novel classes due to the very limited amount of labelled data. Towards reducing prototype biases, we expand the support set using pseudo-labelled UD, where labels are assigned according to respective prediction scores. The prediction scores $f_s(u)$ are again obtained with Eq. 6.7 using updated prototype estimates and current model parameters. We mitigate the varying quality of pseudo labels by selecting the top-K samples with the most confident predictions per class which, by definition, consist of the K unlabelled samples that are closest to their assigned class' prototypes. This augmentation results in an extended annotated support set defined for each class j as $S_j^* = S_j \cup U_j$, where $U_j = \text{top-}K(f_s^j(u))$ is the set of unlabelled samples selected for class j. The prototype is then refined using Eq. 6.9 by replacing S with S^*. Crucially, we

Algorithm 3 Dynamic Prototype Refinement

```
Input: labelled examples $\mathcal{S} = \{S_1, \dots, S_j, S_{C_n}\}$, and unlabelled examples $\mathcal{U}$; Number of novel categories: $C_n$ ; number of iterations M; number of fine-tuning steps R; pre-trained student and teacher model parameters $\theta$,$\theta^T$; weighting parameters $\lambda_{ft}$, $\alpha$.
Output: Prototypes of novel categories $W^{**}$, student model parameters $\theta$;
Prototypes initial inference: $W \leftarrow \{p_1^*, p_2^*, \dots, p_{C_n}^*\}$, calculate $p_j^* \leftarrow P(\mathcal{S}_j)$ by Eq (6.9)
For i = 1 to M :
    Explicit prototype refinement
        $U_j \leftarrow$top-$K(f^j_{t,\theta^T,W^T}(u)), \forall j \in 1, \dots, C_n$,   $f^j_{t,\theta^T,W^T}$ computed as in by Eq (6.7) with parameters $\theta^T, W^T$   ▷ $\theta^T, W^T$ initialised to $\theta, W$ for $i = 1$
        $S_j^* \leftarrow U_j \cup S_j \quad \forall j \in 1, \dots, C_n$
        $W^* \leftarrow \{P(S_1^*), \dots, P(S_{C_n}^*)\}$
    Implicit refinement using CR
    Randomly re-initialise teacher parameters $\theta^T$
    For r = 1 to R:
        Sample a batch of unlabelled samples $\mathcal{U}_s$ from $\mathcal{U}$
        $\bar{u} \leftarrow WeakAugment(u)$, $\hat{u} \leftarrow StrongAugment(u), u \in \mathcal{U}_s$
        $V_j \leftarrow$top-$K(f^j_{t,\theta_t,W*}(\bar{u})) \quad \forall j \in 1, \dots, C_n$
        $W^{**}, \theta^* \leftarrow \underset{W,\theta}{\operatorname{argmin}} \mathcal{L}_{ce}(\mathbf{x}) + \lambda_{ft}\mathcal{L}_{cons}(v_b^u), \quad \mathbf{x} \in \mathcal{S}, v^u \in V = \{V_1, \cdots, V_{C_n}\}$
        Update teacher parameters $W^T = (1-\alpha)W^T + \alpha W^{**}, \theta^T = (1-\alpha)\theta^T + \alpha\theta^*$
end
```

emphasise that per stage pseudo-labels are used *uniquely* to update prototypes and that samples, pseudo-labelled at this stage, are considered unlabelled again at the next iteration. Importantly pseudo-labels are therefore not propagated, allowing recovery from potentially erroneous predictions during the subsequent fine-tuning stage.

(3) Implicit Refinement using Consistency Regularisation. Our implicit refinement stage borrows ideas from gradient-based FSL, which typically adapts the entire model to the novel set of classes via a fine-tuning stage. This stage is generally missing from prototype-based methods, which explicitly represent prototypes as an average feature representation, and thus lose the flexibility afforded by learning implicit network parameters. This fine-tuning stage is particularly desirable in our setting, where we seek to maximally leverage the UD available and where our prototypes are defined as model weights. It is a natural choice to consider deploying

Consistency Regularisation to fine-tune the model, noting that the refined prototypes obtained at this stage afford high quality teacher predictions. We implement the strategy described in Sec. 6.2.1 to fine-tune the model on novel classes with CR. To further improve robustness to noisy teacher predictions and difficult examples, we adopt a selective prototype CR strategy. By calculating teacher prediction scores $f_t(\bar{u})$ according to their prototype distance, we can select the top-K unlabelled examples with the least ambiguous label predictions to compute the CR loss. Note that this second top-K selection set V will differ from top-K set U computed during the explicit stage, as 1) prototypes were updated 2) they are computed on the teacher model subject to weak input augmentation. The model is fine-tuned for R gradient updates by minimising $\mathcal{L}(\mathbf{x}, v_b^u) = \mathcal{L}_{ce}(\mathbf{x}) + \lambda_{ft}\mathcal{L}_{cons}(v^u)$, where $\mathcal{L}_{ce}$ and $\mathcal{L}_{cons}$ are computed as described in Sec. 6.2.1, where labelled sample $\mathbf{x}$ is from $\mathcal{D}_{novel}$ and $v^u \in V$.

Dynamic Refinement. Our implicit and explicit refinement steps allow us to iteratively refine prototypes to further improve performance. As shown in Algorithm 3, we alternate between explicit and implicit steps for M iterations, reinitialising estimated pseudo-label at each iteration. Top-K selection, for the first explicit stage, relies on student predictions since teachers are randomly initialised. Teacher predictions, expected to be more accurate and stable, are used in subsequent iterations. Importantly, we note that teacher parameters are reinitialised before each implicit stage (after explicit selection), so as to introduce stochasticity, increasing robustness to pseudo-label errors and allowing a more diverse exploration of the loss landscape.

6.2.2 Experiments

Experimental set-up. We evaluated Fewmatch on standard SS-FSL benchmarks: *mini*ImageNet [61] and *tiered*ImageNet [183], both subsets of the ImageNet dataset[3] designed for FSL as discussed in Section 6.1. In contrast to fully supervised learning for *mini*ImageNet in Section 6.1, we adopt semi-supervised setting on this benchmark, using 40%/60% of the data for labelled/unlabelled splits following previous works [183, 71]. Similarly, we follow the standard semi-supervised split [183, 71] on *tiered*ImageNet [183], with 10% of the images of each class forming the labelled split and the remaining 90% being the unlabelled data. We consider $K_n^l = 5$ way $N = 1/5$ shot classification problems and follow the strategy adopted in [183, 71] to generate test episodes: we randomly sample K_n^l classes from the test set, N labelled images from each class and 100 unlabelled images as support images, and 15 images as query images. We further test the distractor setting, randomly selecting 100 unlabelled images from 3 task-irrelevant classes as distractors

Table 6.4: Mean classification accuracies of the 5-way 1/5-shot tasks on *mini*ImageNet and *tiered*ImageNet (**Bold**: Best results per set-up).

Setting	Model	Backbone	*mini*Imagenet		*tiered*Imagenet	
cline4-6			1-shot	5-shot	1-shot	5-shot
SL	CTM [255]	*ResNet-18*	62.05 ±0.55	78.63 ±0.06	64.78 ±0.11	81.05 ±0.52
	CCrot[179]	*WRN-28-10*	62.93 ±0.45	79.87 ±0.33	70.53 ±0.51	84.98 ±0.36
	MTL [168]	*ResNet-12*	61.20 ±1.80	75.50 ±0.80	-	-
SSL	MS k-Means[183]	4Conv	50.4	64.4	52.4	69.9
	MS k-Means with MTL	ResNet-12	62.1	73.6	68.6	81.0
	TPN[70]	4Conv	52.8	66.4	55.7	71.0
	TPN with MTL	ResNet-12	62.7	74.2	72.1	83.3
	LST [71]	ResNet-12	70.1 ±1.9	78.7 ±0.8	77.7 ±1.6	85.2 ±0.8
	Ours	ResNet-12	**75.66±0.95**	**82.93±0.62**	**78.70±0.93**	**85.40±0.58**
			Distractor Setting			
SSL	MS k-Means [183] with MTL	ResNet-12	61.0	72.0	66.9	80.2
	TPN [70] with MTL	ResNet-12	61.3	72.4	71.5	82.7
	LST [71]	ResNet-12	64.1	77.4	73.4	83.4
	Ours	ResNet-12	**70.35±0.98**	**80.23±0.66**	**74.24±0.95**	**83.64±0.63**

[71] to be added to the unlabeleld set. We report the mean accuracy of 600 randomly generated test episodes with 95% confidence intervals.

The method was implemented using PyTorch [264]. We use the same backbone architecture ResNet-12 as [71]. For base category training, we follow parameters used in [175]: our model is optimised using SGD with momentum set to 0.9, weight decay to 0.0005, mini-batchsize to 256 (128 LD and 128 UD) for 30 epochs. All input images were resized to 84×84. The learning rate was initialised to 0.1, and updated to 0.01 at epoch 20. Following SSL practice [76], weighting parameter λ is defined as a linear ramp-up function increasing from 0 to 300 in the first 15 epochs. We set the total number of DPR iterations as $M = 3$ and each implicit refinement step fine-tunes the model for 20 steps with 0.01 learning rate. Each mini-batch comprises all LD and 40 randomly sampled UD per-category. We linearly increase weighting parameter λ_{ft} from 0 to 10 in the first 10 steps. The number of unlabelled samples selected is set to $K = 25$. We set EMA rate $\alpha = 0.5$, and $\mathcal{T} = 0.5$. Strong augmentations for the student network are computed using RandAugment [267], applying three random operations with magnitude set to 9. Teacher weak

Table 6.5: Model design analysis on *mini*Imagenet. PCR: base training prototype Consistency Regularisation; ER: Explicit prototype refinment; IR: Implicit refinement using Selective Consistency Regularisation; DR: Dynamic Refinement

Model Components				*mini*ImageNet	
PCR	ER	IR	DR	1-shot	5-shot
Remixmatch				53.52	66.50
Imprinted-weights (IW)				59.09	75.59
IW + Remixmatch (no mixup)				62.20	76.31
✓				61.59	77.90
✓	✓			71.35	81.75
✓	✓	✓		72.52	82.25
✓	✓	✓	✓	75.66	82.93

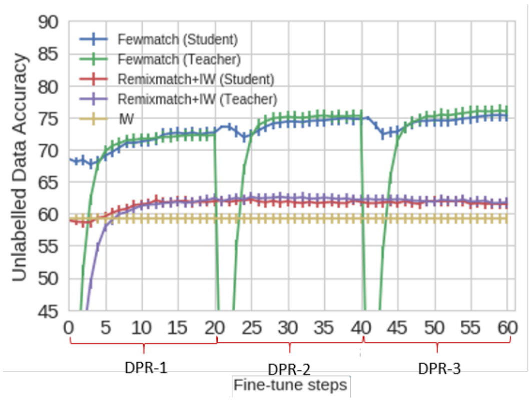


Figure 6.6: Accuracy on training unlabelled data with $M = 3$ iterations of the DPR stage.

augmentations use random cropping and flipping functions.

Comparison to State-of-the-Art (SOTA) Methods. We compared FewMatch with SOTA approaches including (a) 3 FSL and (b) 5 SS-FSL methods in Table 6.4. We note that several SS-FSL approaches, including FewMatch, outperform SOTA FSL approaches, highlighting the potential of using additional UD to learn more accurate models. We observe that FewMatch outperforms the SS-FSL state of the art and that strongest performance gains are observed in the 1-shot setting. We further highlight that the closest SOTA method LST, requires, in contrast to FewMatch, complex episode training, requiring fixed numbers of LD and UD at both training and test time.We additionally compared FewMatch to the SS-FSL SOTA when the UD contains distractor samples. Results are reported in Table 6.4 (distractor setting) and show that we consistently achieve the best performance. This highlights the strong performance and robustness of our method in a more realistic setting.

Ablation experiments. We evaluate the influence of each model component on *mini*ImageNet on 5 way 1/5 shot classification settings. Specifically, we evaluate the influence of using CR in the base training stage (PCR), Explicit Prototype refinement (ER), Implicit Refinement (IR) and Dynamic Refinement (DR) which iterates between ER and IR. We additionally include three baselines: Imprinted Weights[69] (no use of unlabelled data), SOTA CR based SSL method Remixmatch[79] (no accounting for the few-shot setup), and Imprinted Weights combined with Remixmatch. We note that methods using Remixmatch use CR during both base and novel training stage and that the latter method is implemented without mixup (which is used in the Remixmatch method) as the label mixing strategy is not compatible with the prototype approach and would require the definition of infinitely many prototypes. Results are reported in Table 6.5 and show that each component makes a clear contribution to the performance gain; with ER (providing a strong initialisation) and DR (addressing CR's low convergence rates) yielding the strongest performance gains.

Unlabelled data prediction accuracy. Figure 6.6 evaluates the improved reliability of teacher predictions due to our DPR process (M=3). We report accuracy on training UD during the DPR stage, compared to baseline imprinted weights + remixmatch (IWR) which uses CR without addressing the underlying challenges. We observe that our iterative process continuously improves performance, successfully exploiting CR towards reaching higher quality predictions. Conversely, the IWR model fails to exploit UD, obtaining a minimal performance gain with respect to baseline FSL method IW.

6.3 Summary

This chapter provides solutions for cross-task knowledge transfer in image classification. To solve the issue of ineffective cross-task knowledge transfer brought by the single prototype assumption, this chapter introduces a novel, generic approach for metric-based few-shot classification methods. We propose to represent a class as a mixture of prototypes and carefully design our model so as to jointly optimise prototypes variance and representative power. Our model can easily be integrated into a large number of metric learning methods and, as an example, we combined MP with the imprinted weights method yielding performance that significantly outperforms the state of the art on two popular FSL benchmarks. We demonstrate the validity of our design through a comprehensive set of ablation experiments. Existing metric based FSL approaches typically limit class representation to a uni-modal prototype and our work offers a solution to the important limitations commonly associated with such strategies.

Additionally, to take full advantage of unlabelled data in cross-task knowledge transfer, this chapter proposes a novel prototype-driven approach named FewMatch, which exploits, for the first time, the concept of consistency regularisation in an SS-FSL setting. In contrast with pre-existing methods, we alleviate requirements for iterative pseudo-labelling, preventing propagation of errors induced by inaccurate model predictions. Alternatively, we introduce a dynamic prototype refinement strategy that alternates between explicit pseudo label based updates and implicit model fine-tuning. Our extensive experiments demonstrate that this iterative strategy allows successful exploitation of unlabelled data within a consistency regularisation framework.

Chapter 7

Conclusion and Future Work

Sometimes it's necessary to go a long distance out of the way in order to come back a short distance correctly.

—— **Edward Albee**

7.1 Conclusion

This thesis has explored various aspects of Knowledge Transfer (KT) in object recognition, with the goal of yielding more scalable visual recognition systems applicable in real-world scenarios. This is realised by simulating the human behaviour of cross-utilizing the knowledge from the experience, when encountering new complex scenarios. Specifically, four types of knowledge transfer scheme on fine-grained and coarse-grained object recognition task are investigated, including 1) cross-class KT in person re-id; 2) cross-domain KT in person re-id; 3) cross-model KT in image classification; 4) cross-task KT in image classification. These problems are inherently challenging due, in part, to the common convention held by machine learning approaches to ignore knowledge from past experience. This is in direct contrast to humans, and other biological learning systems, capable of displaying genuine intelligence. Crucially, this artificial trait results in lack of generalization abilities, that in turn hinder performance in the complex and novel scenarios that are often met in reality.

In particular,

1. In chapter 3, a novel Identity Discriminative Attention Reinforcement Learning (IDEAL)

framework is proposed to solve the negative influence of background noise to the cross-class knowledge transfer in person re-id. This allows to obtain more accurate and generalizable feature representation of person identity in deployment, and therefore boosting re-id system performance. Besides, this chapter investigates the multi-scale challenge of person bounding box in more severe and realistic situation: person search. To address this limitation, a new Cross-Level Semantic Alignment (CLSA) framework is proposed. The CLSA enables matching images to be "scale-invariant" (more "scale insensitive") in the sense that a scale change in person search.

2. In chapter 4, cross-domain knowledge transfer in person re-id is discussed. To effectively address the domain shift problem when transferring knowledge across domains, an extra local instance alignment constraint is formulated to find the optimal solution for feature network. Furthermore, this chapter discusses more realistic universal person re-id problem by considering the knowledge transfer to any target domains rather than the specific single domain. The proposed Universal Model Learning (UML) for domain-generic universal person re-id is in a "train once, run everywhere" pattern. This differs from all the existing state-of-the-art supervised and unsupervised learning approaches typically considering a "train once, run once" pattern, suffering from per-domain repeated model training as well as the corresponding various costs and limitations.

3. In chapter 5, to address the expensive cost of training teacher models and complex multi-stages in training for cross-model knowledge transfer, self-referenced deep learning (SRDL) and knowledge distillation by On-the-fly Native Ensemble (ONE) are proposed, respectively. In particular, SRDL exploits the self-discovered knowledge in a two-step training procedure without the need of training a large capacity teacher networks. In contrast, ONE constructs a multi-branch variant of a given target network by adding auxiliary branches and creates a native ensemble teacher model from all branches on-the-fly in the single stage training stage. The proposed ONE does not require the complex multi-stage training and reduces expensive cost in training heavy teacher networks, while boosting the model performance through more effectively cross-model knowledge transfer.

4. In chapter 6, cross-task knowledge transfer in image classification is investigated. This chapter focuses on few shot classification, aiming to recognize novel objects with very

limited labelled data. Existing metric based approaches constitute the highly popular strategy to solve few shot classification. However, the common assumption that results in employing a single prototype in current methods can not deal with complex multi-model distribution in reality and results in a model that is very sensitive to the quality of the prototype. A novel mixture of prototypes is formulated in this thesis to address these limitations. Besides, a new semi-supervised few shot classification (SSFSC) method fewmatch is proposed to take full advantage of unlabelled data in cross-task knowledge transfer. In contrast to existing competitive SSFSC approaches based self-training, the proposed fewmatch successfully utilizes the consistency regularization, enabling the fast convergency, avoiding the error propagation and obtaining the state-of-art performance.

7.2 Future Work

The potential research directions beyond the proposed methods in this thesis are summarised as follows:

1. (Chapter 3) **Cross-class knowledge transfer in Person Re-identification:** Chapter 3 attempts to alleviate the negative influence of irrelevant factors to the cross-class knowledge transfer, such as background noise and multi-scale matching challenge. The approaches proposed in Chapter 3 are supervised learning methods, and therefore require expensive human labelling cost. One possible direction could be formulated as unsupervised learning cross-class knowledge transfer, *i.e.*, clustering [232, 117] methods. In this context, the background noise and multi-scale challenge could be further explored.

2. (Chapter 4) **Cross-domain knowledge transfer in person re-identification:** Chapter 4 explores the local instance alignment in cross-domain knowledge transfer and proposes a novel and practical challenge: universal re-id, which is trained in the seed domain and deploys in any unseen target domains. Chapter 4 develops a simple method Univeral Model Learning driven by a variety of data augmentation method. This is the first attempt to solve universal person re-id and might inspire many works. One of the possible directions is considering the UML is an interesting combination of domain generalisation [268] (aiming for target domain performance without data access but no need to maintain the source domain performance) and incremental learning (need to maintain the old class/domain

performance). In this context, some existing gradient based method domain generalisation methods [268, 269] might be useful for universal re-id.

3. (Chapter 5) **Cross-model knowledge transfer in image classification:** Our SRDL and ONE model in chapter 5 investigate the knowledge transfer in human design neural network. Recently, researchers are interested in Neural Architecture Search (NAS) [87, 88, 89, 90, 91, 92], and therefore the combination of NAS and cross-model knowledge transfer is a potential direction. Some people point out the transferred knowledge in NAS contains the structual knowledge of the model [270]. This indicates different achitectures might have an inidividual ability to transfer and receive the knowledge to other network. Thus, how to find a optimal achitecture to easliy learn knowledge from other networks is a open question and worth to be discussed.

4. (Chapter 6) **Cross-task knowledge transfer in image classification:** Although the proposed MP and Fewmatch make a significant step to explore the cross-task knowledge transfer in the more realistic scenarios (complex multi-model distribution and contains unlabel data), there are still a long road to explore the knowledge transfer in the few shot classification. Firstly, the issue of domain shift across task is not deeply investigated. Recent work[271, 272] illustrates the influence of the domain shift largely degrades the model generalization performance across the tasks. Besides, Huang et.al [273] argue current widely used benchmarks Omniglot and miniImageNet are too simple because their class semantics do not vary across episodes, which defeat their intended purpose of evaluating few-shot classification methods. Endeavours on the more challenging META-DATASET [274] benchmark should be encouraged.

Bibliography

[1] Hal Hodson. Deepmind and google: the battle to control artificial intelligence, 2019.

[2] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 2019.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[6] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[10] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv e-prints*, 2016.

[11] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[12] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, June 2019.

[13] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv e-prints*, 2016.

[14] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. *arXiv e-prints*, 2016.

[15] Uday Kamath, John Liu, and James Whitaker. Transfer learning: Domain adaptation. In *Deep Learning for NLP and Speech Recognition*. 2019.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Stateline, Nevada, USA, December 2012.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[18] Ross Girshick. Fast r-cnn. *arXiv e-prints*, 2015.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, Melbourne, Australia, August 2017.

[21] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[22] Mohammad Rastegari and et al Ordonez. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[23] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, Toulon, France, April 2017.

[24] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, San Juan, Puerto Rico, USA, April 2016.

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv e-prints*, 2015.

[26] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, Montreal, Canada., December 2014.

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv e-prints*, 2014.

[28] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[29] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[30] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[31] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv e-prints*, 2017.

[32] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. *arXiv e-prints*, 2017.

[33] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. Salt Lake City, Utah, USA, June 2018.

[34] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[35] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[36] Hanxiao Wang, Xiatian Zhu, Shaogang Gong, and Tao Xiang. Person re-identification in identity regression space. *International Journal of Computer Vision*, 126(12):1288–1310, 2018.

[37] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. *arXiv e-prints*, 2017.

[38] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[39] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[40] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, June 2014.

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[42] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[44] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[45] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv e-prints*, 2017.

[46] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, June 2010.

[47] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, Dundee, UK, September 2011.

[48] Zheng, Zhedong, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv e-prints*, 2017.

[49] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[50] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[51] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, June 2019.

[52] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[53] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *arXiv e-prints*, 2018.

[54] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv e-prints*, 2018.

[55] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, June 2019.

[56] Li, Minxian, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[57] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, June 2019.

[58] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, June 2019.

[59] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshop of European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[60] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*, Vancouver, BC, Canada, April 2018.

[61] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[62] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, Long Beach, California, USA, December 2017.

[63] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[64] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[65] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine learning*, Sydney, Australia, August 2017.

[66] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv e-prints*, 2018.

[67] Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *arXiv e-prints*, 2019.

[68] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv e-prints*, 2018.

[69] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[70] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv e-prints*, 2018.

[71] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[72] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

[73] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. Atlanta, USA, June 2013.

[74] Kelsey R Allen, Evan Shelhamer, Hanul Shin, and Joshua B Tenenbaum. Infinite mixture prototypes for few-shot learning. *arXiv e-prints*, 2019.

[75] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

[76] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, Long Beach, California, USA, December 2017.

[77] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv e-prints*, 2016.

[78] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[79] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv e-prints*, 2019.

[80] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, June 2014.

[81] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[82] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

[83] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, June 2005.

[84] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[85] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *arXiv e-prints*, 2017.

[86] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[87] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv e-prints*, 2018.

[88] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv e-prints*, 2017.

[89] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[90] Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical block-wise neural network architecture generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[91] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv e-prints*, 2018.

[92] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[93] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc Le, and Alex Kurakin. Large-scale evolution of image classifiers. *arXiv e-prints*, 2017.

[94] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, January 2019.

[95] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Efficient multi-objective neural architecture search via lamarckian evolution. *arXiv e-prints*, 2018.

[96] Hector Mendoza, Aaron Klein, Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Towards automatically-tuned neural networks. In *Workshop on Automatic Machine Learning*, 2016.

[97] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial Intelligence and Statistics*, pages 528–536. PMLR, 2017.

[98] Arber Zela, Aaron Klein, Stefan Falkner, and Frank Hutter. Towards automated deep learning: Efficient joint neural architecture and hyperparameter search. *arXiv e-prints*, 2018.

[99] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, June 2012.

[100] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, June 2012.

[101] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012.

[102] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, June 2013.

[103] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, June 2014.

[104] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*. Zurich, Switzerland, September 2014.

[105] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 2016.

[106] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[107] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[108] Shengcai Liao and Stan Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[109] Jiaxin Chen, Zhaoxiang Zhang, and Yunhong Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *Image Processing, IEEE Transactions on*, 2015.

[110] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing*, 2013.

[111] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[112] Hailin Shi, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Yang Yang, and Stan Z Li. Constrained deep metric learning for person re-identification. *arXiv e-prints*, 2015.

[113] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[114] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[115] Kodirov, Elyor, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *British Machine Vision Conference*, Swansea, UK, September 2015.

[116] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[117] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, January 2019.

[118] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. In *British Machine Vision Conference*, Newcastle, UK, September 2018.

[119] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[120] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[121] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, Wales, UK, September 2010.

[122] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Highly efficient regression for scalable person re-identification. In *British Machine Vision Conference*, York, UK, September 2016.

[123] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[124] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. *arXiv e-prints*, 2016.

[125] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv e-prints*, 2016.

[126] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, June 2013.

[127] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[128] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[129] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, Sydney, Australia, October 2013.

[130] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *British Machine Vision Conference*, Nottingham, UK, September 2014.

[131] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *arXiv e-prints*, 2016.

[132] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, Florida, USA, June 2009.

[133] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[134] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[135] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast openworld person re-identification. *IEEE Transactions on Image Processing*, 2017.

[136] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *arXiv e-prints*, 2018.

[137] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[138] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[139] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *arXiv e-prints*, 2016.

[140] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv e-prints*, 2014.

[141] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv e-prints*, 2015.

[142] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[143] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[144] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[145] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[146] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv e-prints*, 2017.

[147] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[148] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards unsupervised open-set person re-identification. In *IEEE International Conference on Image Processing*, 2016.

[149] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised ℓ_1 graph learning. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[150] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[151] Slawomir Bak, Carr, Peter, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[152] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[153] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *IEEE International Conference on Computer Vision*, Seoul, Korea, November 2019.

[154] Hong-Xing Yu, Wu, Ancong, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[155] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[156] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2014.

[157] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *IEEE International Conference on Computer Vision*, Seoul, Korea, November 2019.

[158] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, Marseille, France, October 2008. Springer.

[159] Huanhuan Yu, Menglei Hu, and Songcan Chen. Multi-target unsupervised domain adaptation without exactly shared categories. *arXiv e-prints*, 2018.

[160] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *arXiv e-prints*, 2018.

[161] Cristian Bucilua and et al. Model compression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2006.

[162] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[163] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv e-prints*, 2015.

[164] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research*, 2015.

[165] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[166] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv e-prints*, 2018.

[167] Yaoyao Liu, Qianru Sun, An-An Liu, Yuting Su, Bernt Schiele, and Tat-Seng Chua. Lcc: Learning to customize and combine neural networks for few-shot learning. *arXiv e-prints*, 2019.

[168] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

[169] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

[170] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv e-prints*, 2017.

[171] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv e-prints*, 2017.

[172] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv e-prints*, 2018.

[173] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. Lille, France, 2015.

[174] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[175] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[176] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv e-prints*, 2018.

[177] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[178] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[179] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. *arXiv e-prints*, 2019.

[180] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[181] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv e-prints*, 2019.

[182] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv e-prints*, 2017.

[183] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv e-prints*, 2018.

[184] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 1951.

[185] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996.

[186] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

[187] Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *International Conference on Learning Representations*, Toulon, France, April 2017.

[188] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[189] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 2009.

[190] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, June 2005.

[191] Shaogang Gong, Marco Cristani, Change Loy Chen, and Timothy M. Hospedales. The re-identification challenge. In *Person Re-Identification*. Springer, 2014.

[192] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

[193] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning*, Corvallis, Oregon, USA, June 2007.

[194] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[195] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2), 2009.

[196] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[197] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[198] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[199] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised l1 graph learning. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[200] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[201] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. Sample-specific svm learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[202] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[203] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv e-prints*, 2016.

[204] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv e-prints*, 2014.

[205] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. 2012.

[206] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2015.

[207] Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv e-prints*, 2017.

[208] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA Engineer*, 1984.

[209] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, June 2006.

[210] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[211] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[212] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv e-prints*, 2015.

[213] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017.

[214] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[215] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Convolutional channel features. In *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, October 2015.

[216] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2014.

[217] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, June 2014.

[218] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, June 2013.

[219] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, June 2012.

[220] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[221] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[222] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[223] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[224] Ahmed, Jones, and Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, June 2015.

[225] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

[226] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(Mar), 2012.

[227] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2007.

[228] Çağlar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *The Journal of Machine Learning Research*, 2016.

[229] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv e-prints*, 2015.

[230] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Elsevier, 1989.

[231] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations*, Banff, AB, Canada, April 2014.

[232] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *arXiv e-prints*, 2017.

[233] Albert Henry Munsell. *A color notation*. Munsell color company, 1919.

[234] Munsell color system. `https://en.wikipedia.org/wiki/Munsell_color_system`.

[235] Paul Haeberli and Douglas Voorhies. Image processing by linear interpolation and extrapolation. *IRIS Universe Magazine*, 1994.

[236] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[237] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[238] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *European Conference on Computer Vision*, Munich, Germany, September 2018.

[239] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv e-prints*, 2018.

[240] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. 2018.

[241] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 1999.

[242] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine learning*, Bellevue, Washington, USA, June 2011.

[243] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. *arXiv e-prints*, 2017.

[244] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations*, April 2017.

[245] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[246] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and

Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[247] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[248] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[249] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[250] et al. Keskar, Nitish Shirish. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv e-prints*, 2016.

[251] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 2016.

[252] Shen Li Sun Gang Hu, Jie. Squeeze-and-excitation networks. *arXiv e-prints*, 2017.

[253] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. Boosting supervision with self-supervision for few-shot learning. *arXiv e-prints*, 2019.

[254] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

[255] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

[256] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv e-prints*, 2019.

[257] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2018.

[258] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. Long Beach, California, USA, June 2019.

[259] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. *arXiv e-prints*, 2019.

[260] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. *arXiv e-prints*, 2019.

[261] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. *arXiv e-prints*, 2019.

[262] Xiang Jiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, and Stan Matwin. Learning to learn with conditional class dependencies. 2018.

[263] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[264] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[265] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, Montreal, Canada, December 2018.

[266] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[267] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv e-prints*, 2019.

[268] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[269] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv e-prints*, 2018.

[270] Yu Liu, Xuhui Jia, Mingxing Tan, Raviteja Vemulapalli, Yukun Zhu, Bradley Green, and Xiaogang Wang. Search to distill: Pearls are everywhere but not the eyes. In *IEEE Conference on Computer Vision and Pattern Recognition*, Online, June 2020.

[271] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, Ji-Rong Wen, and Ping Luo. Domain-adaptive few-shot learning. *arXiv e-prints*, 2020.

[272] Doyen Sahoo, Hung Le, Chenghao Liu, and Steven CH Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift. *International Conference on Learning Representations*, April 2018.

[273] Gabriel Huang, Hugo Larochelle, and Simon Lacoste-Julien. Are few-shot learning benchmarks too simple? *arXiv e-prints*, 2019.

[274] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv e-prints*, 2019.