# Visual Learning in Limited-Label Regime

**Yanbei Chen**

School of Electronic Engineering and Computer Science

Queen Mary University of London

2021

*Dedicated to my dear parents.*

# Visual Learning in Limited-Label Regime

**Yanbei Chen**

## Abstract

Deep learning algorithms and architectures have greatly advanced the state-of-the-art in a wide variety of computer vision tasks, such as object recognition and image retrieval. To achieve human- or even super-human-level performance in most visual recognition tasks, large collections of labelled data are generally required to formulate meaningful supervision signals for model training. The standard supervised learning paradigm, however, is undesired in several perspectives. First, constructing large-scale labelled datasets not only requires exhaustive manual annotation efforts, but may also be legally prohibited. Second, deep neural networks trained with full label supervision upon a limited amount of labelled data are weak at generalising to new unseen data captured from a different data distribution. This thesis targets at solving the critical problem of lacking sufficient label annotations in deep learning. More specifically, we investigate four different deep learning paradigms in limited-label regime, including *close-set semi-supervised learning*, *open-set semi-supervised learning*, *open-set cross-domain learning*, and *unsupervised learning*. The former two paradigms are explored in visual classification, which aims to recognise different categories in the images; while the latter two paradigms are studied in visual search – particularly in person re-identification – which targets at discriminating different but similar persons in a finer-grained manner and can be extended to the discrimination of other objects of high visual similarities. We detail our studies of these paradigms as follows.

**Chapter 3**: **Close-Set Semi-Supervised Learning** (Figure 1 **(I)**) is a fundamental semi-supervised learning paradigm that aims to learn from a small set of labelled data and a large set of unlabelled data, where the two sets are assumed to lie in the *same label space*. To address this problem, existing semi-supervised deep learning methods often rely on the up-to-date "network-in-training" to formulate the semi-supervised learning objective, which ignores both the disriminative feature representation and the model inference uncertainty revealed by the network in the preceding learning iterations, referred to as the memory of model learning. In this work, we proposed to augment the deep neural network with a lightweight *memory mechanism* [Chen et al., 2018b], which captures the *underlying manifold structure* of the labelled data at the per-class level, and further imposes auxiliary *unsupervised constraints* to fit the unlabelled data towards the underlying manifolds. This work established a simple yet efficient close-set semi-supervised deep learning scheme to boost model generalisation in *visual classification* by learning from sparsely labelled data and abundant unlabelled data.

**Chapter 4**: **Open-Set Semi-Supervised Learning** (Figure 1 **(II)**) further explores the potential of learning from abundant *noisy* unlabelled data, While existing SSL methods artificially assume that small labelled data and large unlabelled data are drawn from the same class distribution, we consider a more realistic and uncurated open-set semi-supervised learning paradigm. Considering visual data is always growing in many visual recognition tasks, it is therefore implausible to pre-define a fixed label space for the unlabelled data in advance. To investigate this new chal-

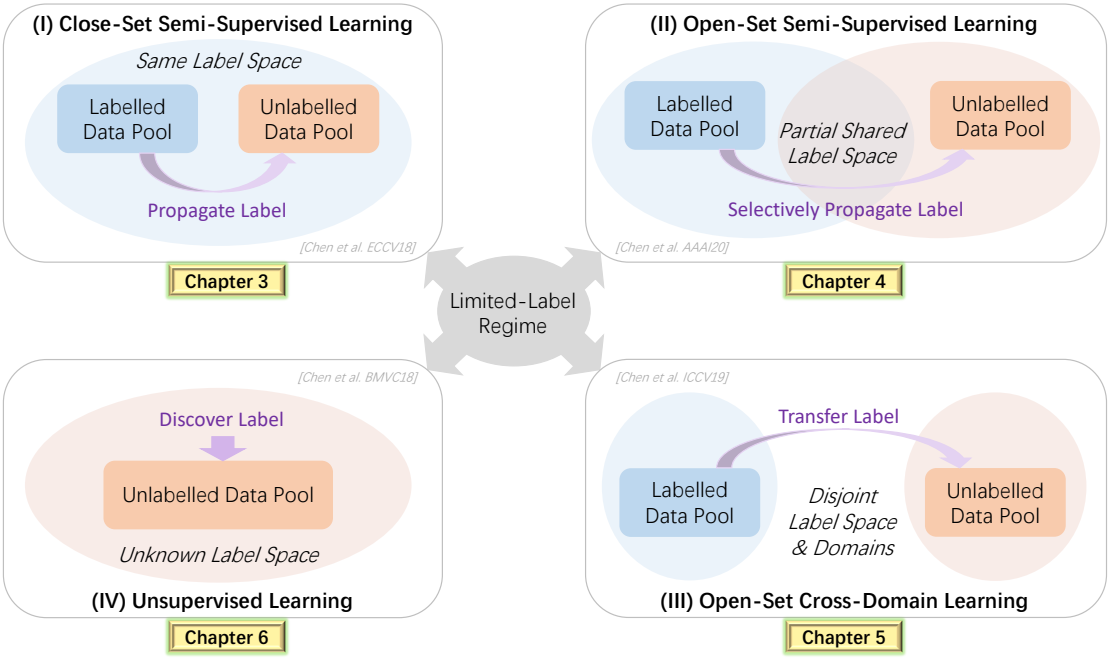



Figure 1: An overview of the main studies in this thesis, which covers four different deep learning paradigms in the limited-label regime, including (I) *close-set semi-supervised learning* (Chapter 3), (II) *open-set semi-supervised learning* (Chapter 4), (III) *open-set cross-domain learning* (Chapter 5), and (IV) *unsupervised learning* (Chapter 6). Each chapter studies a specific deep learning paradigm that requires to *propagate*, *selectively propagate*, *transfer*, or *discover* label information for model optimisation, so as to minimise the manual efforts for label annotations. While the former two paradigms focus on semi-supervised learning for *visual classification*, i.e. recognising different visual categories; the latter two paradigms focus on semi-supervised and unsupervised learning for *visual search*, i.e. discriminating different instances such as persons.

lenging learning paradigm, we established the first systematic work to tackle the open-set semi-supervised learning problem in *visual classification* by a novel approach: *uncertainty-aware self-distillation* [Chen et al., 2020b], which selectively propagates the soft label assignments on the unlabelled visual data for model optimisation. Built upon an *accumulative ensembling* strategy, our approach can jointly capture the model uncertainty to discard *out-of-distribution* samples, and propagate less *overconfident* label assignments on the unlabelled data to avoid catastrophic error propagation. As one of the pioneers to explore this learning paradigm, this work opens up new avenues for research in more realistic semi-supervised learning scenarios.

**Chapter 5**: **Open-Set Cross-Domain Learning** (Figure 1 **(III)**) is a challenging semi-supervised learning paradigm of great practical value. When training a visual recognition model in an operating visual environment (i.e. *source domain*, such as the laboratory, simulation, or known scene), and then deploying it to unknown real-world scenes (i.e. *target domain*), it is likely that the model would fail to generalise well in the unseen visual target domain, especially when the target domain data comes from a *disjoint label space* with heterogeneous *domain drift*. Unlike prior works in domain adaptation that mostly consider a shared label space across two domains, we studied the more demanding *open-set domain adaptation* problem, where both label spaces and domains are disjoint across the labelled and unlabelled datasets. To learn from these heterogeneous datasets, we designed a novel *domain context rendering* scheme for open-set cross-domain learning in visual search [Chen et al., 2019a] – particularly for person re-identification, i.e. a realistic testbed to evaluate the representational power of fine-grained discrimination among very similar instances. Our key idea is to transfer the source identity labels into diverse target domain

contexts. Our approach enables the generation of an abundant amount of synthetic training data that selectively blend label information from source domain and context information from target domain. By training upon such synthetic data, our model can learn a more identity-discriminative and context-invariant representation for effective *visual search* in the target domain. This work sets a new state-of-the-art in cross-domain person re-identification and provides a novel and generic solution for open-set domain adaptation.

**Chapter 6**: **Unsupervised Learning** (Figure 1 **(IV)**) considers the learning scenario with none labelled data. In this work, we explore unsupervised learning in visual search, particularly for person re-identification, a realistic testbed to study unsupervised learning, where person identity labels are generally very difficult to acquire over a wide surveillance space [Chen et al., 2018a]. In contrast to existing methods in person re-identification that requires exhaustive manual efforts for labelling cross-view pairwise data, we aims to learn visual representations without using any manual labels. Our generic rationale is to formulate auxiliary supervision signals that learn to uncover the underlying data distribution, consequently grouping the visual data in a meaningful and structural way. To learn from the unlabelled data in a fully unsupervised manner, we proposed a novel *deep association learning* scheme to uncover the underlying data-to-data association. Specifically, two *unsupervised constraints – temporal consistency* and *cycle consistency* – are formulated upon *neighbourhood consistency* to progressively associate visual features *within* and *across* video sequences of tracked persons. This work sets the new state-of-the-art in video-based unsupervised person re-identification and advances the automatic exploitation of video data in real-world surveillance.

In summary, the goal of all these studies is to build efficient and scalable visual learning models in the limited-label regime, which empower to learn more powerful and reliable representations from complex unlabelled visual data and consequently learn more powerful visual representations to facilitate better visual recognition and visual search.

School of Electronic Engineering and Computer Science

Queen Mary University of London

2021

# Publications

The following publications contain the final reports of the findings in this thesis.

- **Chapter 3**

  **Yanbei Chen**, Xiatian Zhu, Shaogang Gong. *Semi-supervised deep learning with memory* [Chen et al., 2018b]. In European Conference on Computer Vision, Munich, Germany, September 2018. **(ECCV)**

- **Chapter 4**

  **Yanbei Chen**, Xiatian Zhu, Wei Li, Shaogang Gong. *Semi-supervised learning under class distribution mismatch* [Chen et al., 2020b]. In Association for the Advancement of Artificial Intelligence, New York City, USA, February 2020. **(AAAI)**

- **Chapter 5**

  **Yanbei Chen**, Xiatian Zhu, Shaogang Gong. *Instance-guided context rendering for cross-domain person re-identification* [Chen et al., 2019a]. In International Conference on Computer Vision, Seoul, Korea, October 2019. **(ICCV)**

- **Chapter 6**

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

**Publications in relevant topics were also made during the course of this thesis:**

1. **Yanbei Chen**, Shaogang Gong, Loris Bazzani. *Image Search with Text Feedback by Visiolinguistic Attention Learning* [Chen et al., 2020a]. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, June 2020. **(CVPR)**

2. **Yanbei Chen**, Loris Bazzani. *Learning Joint Visual Semantic Matching Embeddings for Language-Guided Retrieval* [Chen and Bazzani]. In European Conference on Computer Vision, Online, August 2020. **(ECCV)**

3. **Yanbei Chen**, Xiatian Zhu, Shaogang Gong. *Person Re-Identification by Deep Learning Multi-Scale Representations* [Chen et al., 2017c]. In International Conference on Computer Vision, Workshop on Cross-Domain Human Identification, Venice, Italy, October 2017. **(ICCVW)**

## 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. The works mentioned above have all been published.

# Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Shaogang Gong for his long, persistent guidance and enthusiastic encouragement during my PhD study. Meanwhile, I want to convey my sincere thanks to Dr. Xiatian Zhu, a wonderful collaborator who offers uncountable help and invaluable advices through my whole PhD process. It is through their inspiring supervision that I learn to drive research ideas independently and creatively.

I would also like to thank Dr. Miles Hansard and Dr. Qianni Zhang for being my progression panel during my PhD study. My special appreciation goes to all of my colleagues at the Vision Group for their friendship and support in the past years: Zhiyi Cheng, Jingya Wang, Qi Dong, Jiabo Huang, Guile Wu, Pan Li, Xu Lan, Wei Li, Hanxiao Wang, Aytac Kanaci, Qingze Yin, Hang Su, Minxian Li, Zimo Liu, Li Zhang, Weihong Li, Guan'an Wang. I am very grateful for all the nice and supportive QMUL system officers for their generous help in configuring the computing facilities. My special thanks go to Tim Kay, who has been devoted to help generously in a timely manner.

Finally, I would like to thank my parents and boyfriend. Their unconditional love, care, support, and understanding have been serving as a solid pillar in my PhD life.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

---

Deep learning algorithms and architectures [LeCun et al., 2015; Goodfellow et al., 2016] have greatly advanced the state-of-the-art in the computer vision community. Built upon multiple levels of non-linear mathematic operations, deep models can learn meaningful data representations in an increasingly order of abstraction, by training upon a tremendous amount of annotated visual data. The powerful representation learning capability has greatly altered the precedent mindset of engineering hand-crafted visual features. Nowadays, deep learning has replaced feature engineering, notably providing more successful solutions to tackle all sorts of computer vision tasks, such as object recognition [Krizhevsky et al., 2012], retrieval [Schroff et al., 2015], detection [Ren et al., 2015], and segmentation [Chen et al., 2017a].

However, to achieve humanlike or even superhuman performance in most computer vision tasks, the first and foremost preparation is to collect a large volume of visual data with rich label annotations for supervised training through gradient descent [Rumelhart et al., 1985]. Although supervised deep learning paradigm is effective at absorbing and memorising visual data at a scale that is much larger, faster and better than human, it is undesired in several aspects: First, constructing large-scale labelled datasets requires prohibitively expensive manual annotation efforts; Second, deep neural networks trained with limited labelled training data do not scale and generalise well to test data collected from an unseen distribution.

In this thesis, we research various deep learning paradigms in the limited-label regime – when label annotations are very sparse or even unavailable in the training set. As unlabelled training data could exhibit different properties subjected to distributional discrepancy or the unavailabil-

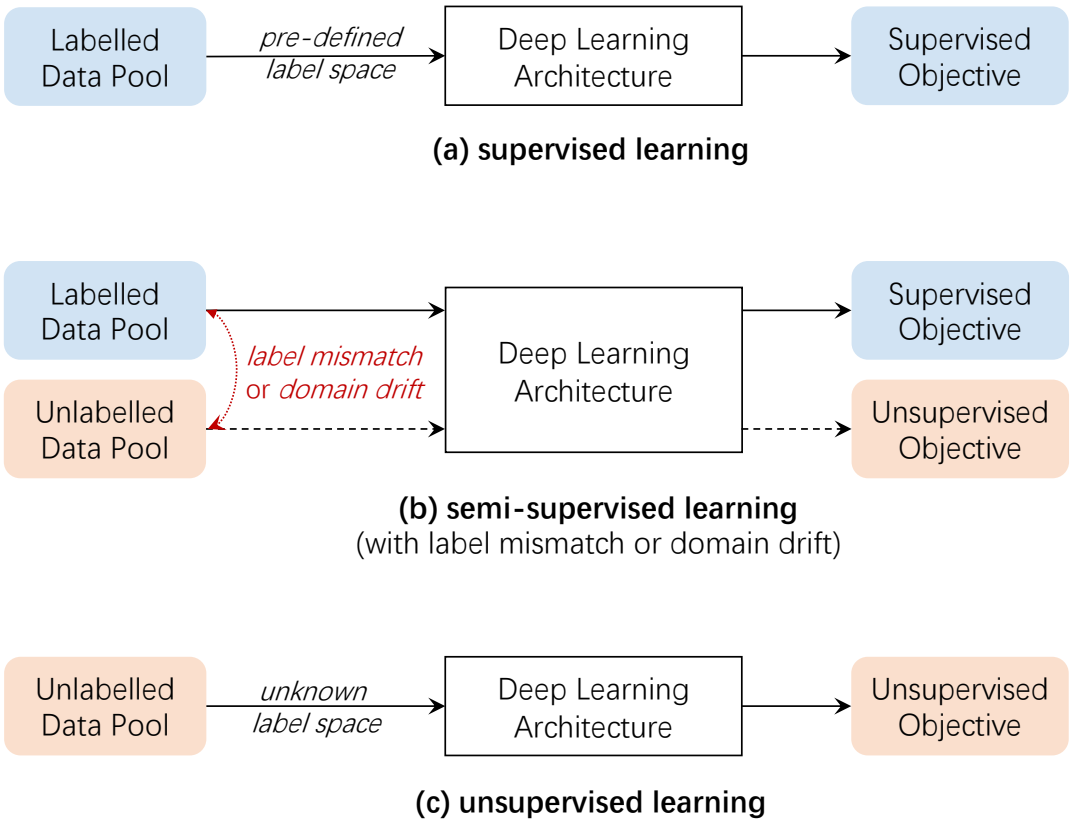



Figure 1.1: An illustration of *supervised*, *semi-supervised* and *unsupervised* deep learning.

ity of labelled data, we study four different learning paradigms that consider unlabelled data are provided under different conditions, where there exist *label mismatch* and (or) *domain drift* between the labelled and unlabelled sets, or none labelled data is available. As opposed to *supervised* learning that requires large collections of labelled data to formulate *supervised objective* for model optimisation (Figure 1.1 (a)), *semi-supervised* and *unsupervised* learning (Figure 1.1 (b)(c)) could alleviate the requirement of exhaustive manual annotations by learning from unlabelled data based on *unsupervised objective* to fit the model towards the underlying structures in visual data. Undoubtedly, the latter paradigms are more promising to cope with the complexity and uncertainty of the fast-growing collections of visual data in this ever-changing digital era.

To explore semi-supervised and unsupervised learning paradigms, we target at two basic and critical computer vision tasks, namely *visual classification* and *visual search*, both of which are especially driven by effective visual representation learning. Before delving into our formulated deep learning paradigms to tackle these two tasks in Section 1.3, we provide basic problem definitions in Section 1.1 and introduce our application domains in Section 1.2.

## 1.1 Basic Problem Definitions

### 1.1.1 Supervised, Semi-Supervised and Unsupervised Learning

As depicted in Figure 1.1, deep learning algorithms can be grouped into three classes depending on the availability and amount of label annotations in the training set, as defined below.

1. **Supervised Learning** is the most typical deep learning paradigm that learns a function $f(\cdot)$ to predict the posterior distribution over a pre-defined set of class label of the labelled data $\mathbf{x}$: $P(\mathbf{y}|\mathbf{x}) = f(\mathbf{x})$; while the goal is to ensures samples of the same class have similar predictive distributions and representations.
2. **Semi-Supervised Learning** considers to estimate the posterior distribution $P(\mathbf{y}|\mathbf{x}) = f(\mathbf{x})$ from the collections of labelled data $P(\mathbf{x}_l, \mathbf{y}_l)$ and unlabelled data $P(\mathbf{x}_u, \mathbf{y}_u)$. The goal is also to ensure samples of the same class to have similar representations. However, additional complexity and uncertainty can be induced if there exist *label mismatch* and (or) *domain drift* between the labelled data and unlabelled data, i.e. $P(\mathbf{x}_l, \mathbf{y}_l) \neq P(\mathbf{x}_u, \mathbf{y}_u)$.
3. **Unsupervised Learning** does not utilise any task-related labelled data, but instead learns a representation $\mathbf{h} = f(\mathbf{x})$ that groups unlabelled samples in a meaningful manner.

As various possible types of distributional discrepancy – e.g. *label mismatch* and (or) *domain drift* – may be presented between the labelled data and unlabelled data, semi-supervised learning (SSL) can be further characterised into different categories, such as *close-set* and *open-set* learning, *single-domain* and *cross-domain* learning, which are detailed in the following.

### 1.1.2 Close-Set and Open-Set Learning

*Open-set* recognition is considered as opposed to *close-set* recognition, which account for *unknown* classes during testing [Scheirer et al., 2012]. In semi-supervised learning, open-set versus close-set learning can be defined based on whether there exists *label mismatch* between the labelled and unlabelled training data. If the label space of labelled and unlabelled datasets are denoted as $\mathcal{Y}_l$ and $\mathcal{Y}_u$, *close-set* and *open-set* learning can be further defined as below.

1. **Close-Set Learning** considers the unlabelled training data contains only in-distribution samples that lie in the known label space, i.e. $\mathcal{Y}_u \subseteq \mathcal{Y}_l$, as Figure 1.2 (a) shows.
2. **Open-Set Learning** considers the unlabelled training data contains out-of-distribution

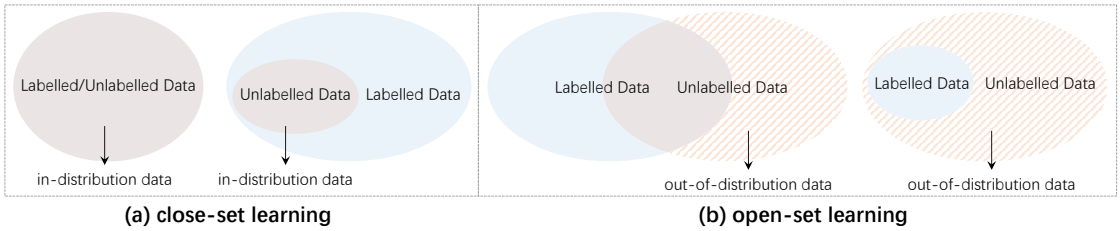



Figure 1.2: An illustration of *close-set* and *open-set* learning.

samples that do not belong to any of the known classes, i.e. $\mathcal{Y}_u - \mathcal{Y}_u \cap \mathcal{Y}_l \neq \varnothing$, as Figure 1.2 (b) shows.

In brief, based on the distributions of label space for labelled data and unlabelled data, one can easily categorise semi-supervised learning as close-set or open-set. In particular, in open-set semi-supervised learning, there exists class distribution mismatch between the labelled and unlabelled data, which is a more challenging learning paradigm as the unlabelled data is contaminated with out-of-distribution samples from unknown classes.

### 1.1.3 Single-Domain and Cross-Domain Learning

Besides *label mismatch*, *domain drift* can also exist between the labelled and unlabelled sets, which further categorise semi-supervised learning into with-domain and cross-domain. In essence, cross-domain semi-supervised learning is one typical scenario in domain adaptation, where the goal is to tackle the intrinsic distributional drift presented across different visual domains [Saenko et al., 2010; Ganin and Lempitsky, 2015]. Such distributional drift could be commonly caused by the cross-domain discrepancy in illumination, viewpoint, resolution, occlusion, and background clutter induced by different scenes, camera characteristics, times of the day, and weather conditions. When incorporating additional unlabelled data for semi-supervised learning, it is therefore essential to consider the potential *domain drift* between the labelled and unlabelled datasets. If the data distributions of the source labelled data and target unlabelled data are written as $P(X)_l^s$ and $P(X)_u^t$, *single-domain* and *cross-domain* learning can be defined as below.

1. **Single-Domain Learning** considers the labelled and unlabelled training data share the same underlying data distribution, i.e. $P(X)_l^s = P(X)_u^t$.
2. **Cross-Domain Learning** considers the labelled and unlabelled training sets are drawn from different visual domains, with unmatched data distributions, i.e. $P(X)_l^s \neq P(X)_u^t$.

To study different challenges of visual learning in limited-label regime, we consider multiple types of semi-supervised learning paradigms that may involve *label mismatch* and *domain drift*. We also consider the case of lacking any label annotations in the training data, i.e. unsupervised learning. These studied paradigms are further detailed in Section 1.3, which all together present a comprehensive research exploration on how to effectively exploit unlabelled data under different conditions to learn more generic visual representations.

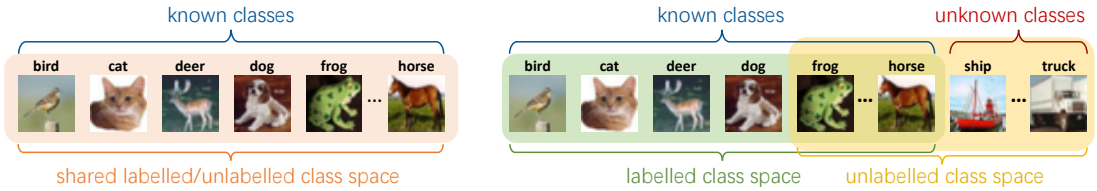
## 1.2 Semi-Supervised and Unsupervised Visual Representation Learning

One of the essential groundings to enhance model performance in computer vision tasks lie in learning more discriminative representations (a.k.a. features) from the visual data. Rather than rely on human knowledge to design hand-crafted visual features, deep learning methods automate the feature engineering process by learning representations from a tremendous amount of visual data. This is achieved by a combination of multiple factors, including designing advanced network architectures [Krizhevsky et al., 2012; He et al., 2016], formulating discriminative learning constraints [Schroff et al., 2015; Wen et al., 2016], introducing auxiliary regularisation techniques [Krogh and Hertz, 1992; Ioffe and Szegedy, 2015], and training with effective optimisers [Tieleman and Hinton, 2012; Kingma and Ba, 2014]. To further improve model generalisation in more complex real-world visual environments, another promising way is to learn from auxiliary unlabelled dataset by semi-supervised learning or unsupervised pre-training. By formulating meaningful unsupervised learning objectives upon the unlabelled training data, more generic visual representations can be learned to extract the discriminative visual information.

In this thesis, we study semi-supervised and unsupervised visual learning mainly in two critical vision application domains, including visual classification and visual search, which are further introduced below to analyse their underlying algorithmic challenges.

### 1.2.1 Application Domain I: Semi-Supervised Visual Classification

Visual classification is one of the most fundamental tasks in computer vision. Over the last years, deep learning has significantly advanced the field of visual classification, achieving or even surpassing human-level performance on various benchmark datasets ranging from simple digit classification (e.g. MNIST [LeCun et al., 2010], SVHN [Netzer et al., 2011]), to more complex natural object recognition (e.g. CIFAR [Krizhevsky and Hinton, 2009], ImageNet [Deng




(a) close-set semi-supervised learning (b) open-set semi-supervised learning

Figure 1.3: An illustration of **(a)** *close-set* and **(b)** *open-set* semi-supervised learning paradigms in **visual classification**.

et al., 2009]) and fine-grained visual categorisation (e.g. Stanford Dogs [Khosla et al., 2011], Caltech UCSD Birds [Wah et al., 2011]). The task of visual classification can be categorised as *single-label* or *multi-label* classification, i.e. tagging each image with one or multiple labels, which requires *one* or *multiple* label annotations per sample for supervised learning. Accordingly, human-level performance of deep learning algorithms generally relies heavily on exhaustive human supervision, i.e. manually annotating the visual data at a sufficient large data scale.

**Semi-Supervised Visual Classification.** The requirement of tremendous label annotations, however, could greatly restrict the generalisation of deep learning models, when diverse visual data with rich label annotations are very expensive or even prohibitive to acquire. Hence, it is non-trivial to formulate semi-supervised learning paradigm [Chapelle et al., 2009] that can exploit both labelled and unlabelled visual data to improve model generalisation. In general, semi-supervised learning objective should fulfil two goals jointly: (1) classifying labelled training data correctly; while (2) learning to classify the auxiliary unlabelled training data.

**Challenges.** Algorithmically, semi-supervised visual classification faces a number of challenges. First, sufficient label annotations are lacking to train a discriminative classifier that could correctly categorise either unlabelled data or test data. Second, class distribution mismatch may exist between the labelled and unlabelled data, which indicates the noisy and task-unrelated unlabelled data could possibly degrade the model performance.

**Task Description.** Herein, as illustrated in Figure 1.3, to research on solutions for the aforementioned challenges, we study a standard *close-set* semi-supervised learning paradigm and a new under-explored *open-set* semi-supervised learning paradigm in visual classification – while the former merely considers to tackle the aforementioned first challenge, the latter tackles both of the challenges jointly. Both learning paradigms are further outlined in Section 1.3.

### 1.2.2 Application Domain II: Semi-Supervised and Unsupervised Visual Search

Visual search is another fundamental task in computer vision, which brings great practical impact to many real-world applications, such as image search on internet [Weyand et al., 2020], product search in e-commerce [Zhang et al., 2018c; Chen et al., 2020a], face recognition [Schroff et al., 2015], and person re-identification [Gong et al., 2014] in visual surveillance. Unlike visual classification that focuses on discriminating a fixed set of visual categories, visual search requires more fine-grained discrimination between instances, e.g. distinguishing different person identities across non-overlapping cameras distributed over open surveillance space and time. This requires to learn a fine-grained visual representation that is both discriminative to task-relevant essential factors while invariant to task-irrelevant redundant variations.

**Semi-Supervised and Unsupervised Visual Search.** In visual search, it is more challenging to acquire fine-grained label annotations at per-instance level. For example, unlike annotating common objects (e.g. cat and dog) in visual classification that generally relies on common sense, in order to acquire person identity labels for a large-scale population in visual search, human annotators have to memorise all the possible unfamiliar person identities over a wide public space. This is not only intellectually challenging for annotators, but may also introduce noisy annotations due to inevitable human errors. Therefore, it is necessary to introduce semi-supervised and unsupervised learning paradigms that could properly address the bottleneck of lacking sufficient qualified label annotations in visual search.

**Challenges.** As compared to visual classification, visual search generally does not need to pre-define a fixed class label space for model training. In fact, at test time, the main goal is to match the same or similar instances of novel unseen classes in new domain context, rather than categorise instances of seen classes. This raises several unique challenges in semi-supervised or unsupervised visual search. First, the class spaces of labelled and unlabelled datasets can be totally *disjoint*, since unseen visual data generally belongs to novel classes in visual search. Second, there may also exist inevitable *domain gap* between the labelled and unlabelled datasets collected from different environments. Third, it is sometimes prohibited to collect annotations, due to high annotations costs and legislation on data privacy.

**Task Description.** To tackle the lack of label annotations in visual search, we consider *person re-identification* as a special case study (Figure 1.4 **(a)**), where the goal is to retrieve images or videos in database to discover the right person in query image or video. In particular, we study

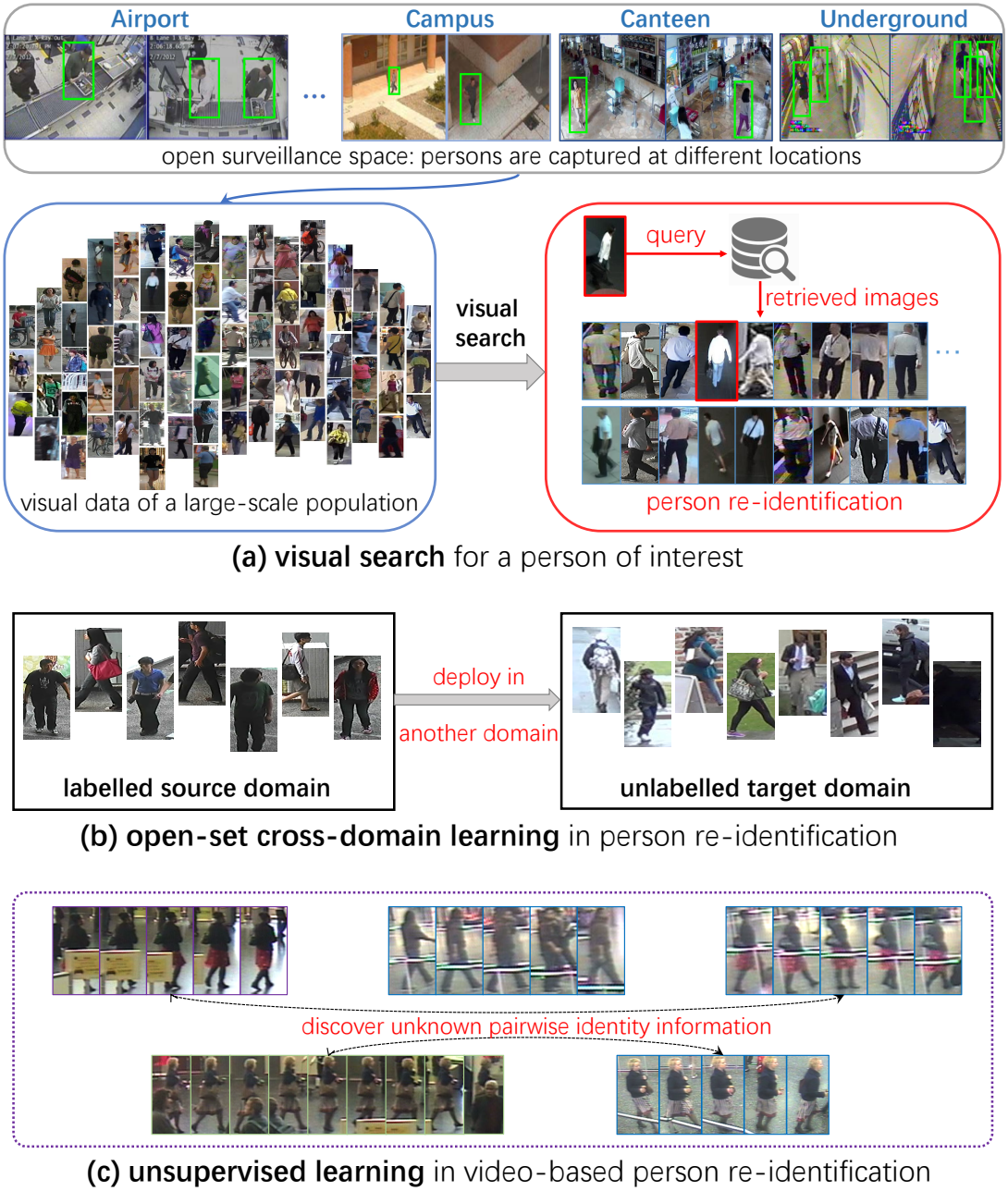



Figure 1.4: An illustration on **(a) visual search** task: *person re-identification*; our studied **(b)** *open-set cross-domain learning*, and **(c)** *unsupervised learning* paradigms in the task of person re-identification.

an *open-set cross-domain learning* paradigm, which addresses the aforementioned challenges of learning from auxiliary unlabelled data that lie in the *disjoint label space* and *another domain* (Figure 1.4 **(b)**). We also study an *unsupervised learning* paradigm that targets at learning discriminative visual representations without utilising any label annotations (Figure 1.4 **(c)**). Both learning paradigms are further outlined in Section 1.3.

## 1.3 Contributions and Thesis Outline

All our studied visual learning paradigms share the same goal of learning from unlabelled data in limited-label regime with minimal human supervision – where little or none labelled data is provided. As unlabelled visual data may emerge differently in different application scenarios, it is imperative to uncover the underlying data structure and mine reliable label information for training. To this end, we introduce various auxiliary supervision signals to propagate, selectively propagate, transfer, and discover label information to learn from the unlabelled visual data. In the sequel, contributions on different learning paradigms are further outlined and summarised.

**[Chapter 3]: Contribution on Close-Set Semi-Supervised Learning**. We tackle the close-set semi-supervised learning scenario in *visual classification*, where a small set of labelled data and a large set of unlabelled data are assumed to share the same class label space. To address the challenge of propagating label assignments reliably to the unlabelled data, we formulate a novel lightweight *memory module* into the network training process [Chen et al., 2018b]. During training, it captures the *underlying manifold structure* of labelled data globally, and imposes *unsupervised learning constraints* that encourage the model to fit the unlabelled data towards the underlying manifold. This work contributes a simple yet efficient semi-supervised deep learning scheme to boost model generalisation by learning effectively from abundant unlabelled data.

**[Chapter 4]: Contribution on Open-Set Semi-Supervised Learning**. We further explore the potential of learning from unconstrained unlabelled data without imposing any *de facto* assumption on the class label space. In particular, we are the first to systematically study and address this new challenging learning paradigm with a comprehensive benchmark and a novel deep learning solution in *visual classification* [Chen et al., 2020b]. We proposed an *uncertainty-aware self-distillation* formulation to selectively propagate soft label assignments on the unlabelled data for model optimisation. Our model formulation can jointly capture the model uncertainty to discard out-of-distribution samples with low confidence scores, and derive less *overconfident* label assignments on the unlabelled data to avoid catastrophic error propagation. This work opens up new avenues for research in more realistic semi-supervised learning scenarios.

**[Chapter 5]: Contribution on Open-Set Cross-Domain Learning**. To avoid the need of collecting a new labelled dataset in an unseen environmental domain, we study an open-set cross-domain learning paradigm in person re-identification, a realistic testbed in visual search that requires to tackle the problem of *label mismatch* and *domain drift* across the labelled and unla-

belled data. Specifically, we designed a novel *domain context rendering* scheme for the task of cross-domain person re-identification [Chen et al., 2019a]. Built with a dual conditional GAN framework, our approach is capable of rendering the source identity population into a diverse range of target domain contexts, consequently leading to a substantial amount of synthetic training data that cover *rich contextual variations* in the target domain environments. By training upon these synthetic data, our model becomes both discriminative to person identities, and more invariant towards task-irrelevant visual variations in the target domain. This work sets the new state-of-the-art in cross-domain person re-identification, and provides a generic solution in open-set domain adaptation.

**[Chapter 6]: Contribution on Unsupervised Learning**. To learn good visual data representations without any manual labels, the key challenge is to formulate reliable auxiliary supervision signals that could uncover the underlying data distribution and group the unlabelled visual data in a meaningful and structural way. To tackle this challenge, we target at learning a more discriminative visual representation in an unsupervised manner for the task of person re-identification [Chen et al., 2018a], which is a realistic testbed to study unsupervised visual search, as large-scale person identity labels are generally very difficult and possibly legally prohibitive to collect. Specifically, we formulated two novel unsupervised learning constraints: *temporal consistency* and *cycle consistency*, both of which incrementally associate the video sequences *within* and *across* camera views to learn discriminative visual representations. The key insight of our approach is to exploit the *neighbourhood consistency* that could progressively uncover the underlying data-to-data association. This work sets the new state-of-the-art in unsupervised person re-identification, and further advances the scalability and applicability of deep networks in automatic exploitation of unlabelled data.

**[Chapter 7]** To conclude this thesis, we further provide a conclusion on semi-supervised and unsupervised visual representation learning in limited-label regime, and discuss other possibilities of introducing auxiliary supervision to improve model generalisation. An outline of all chapters is shown in Figure 1.5.

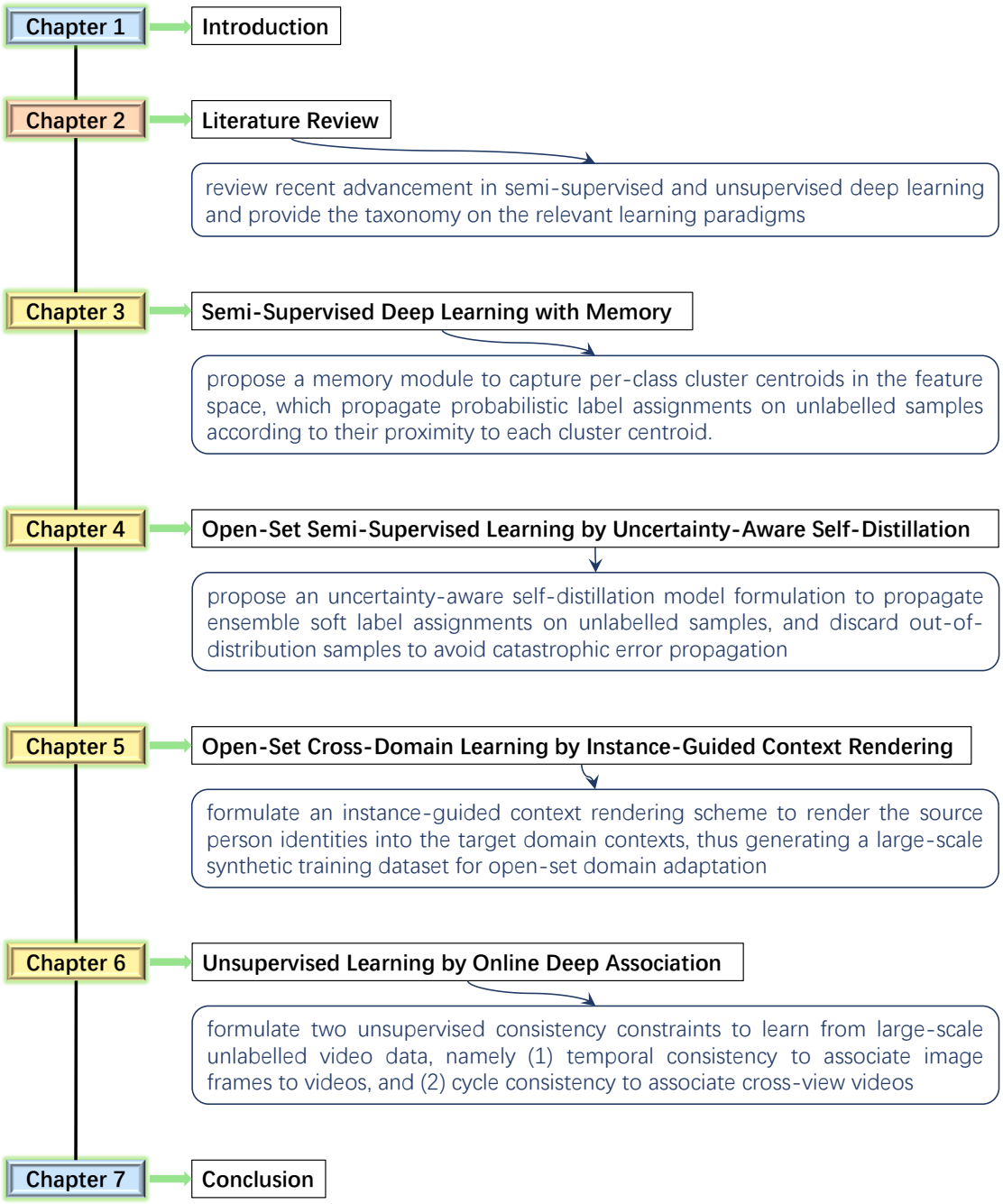



Figure 1.5: An outline of all chapters.

# Chapter 2

# Literature Review

---

In this chapter, we review the relevant developments that target at building visual recognition models without exhaustive human supervision being involved. Specifically, we consider four different visual learning paradigms in the limited-label regime, as defined next.

1. **Close-Set Semi-Supervised Learning** aims to learn from sparsely labelled data and large sized auxiliary unlabelled data, where the two sets are assumed to share the *identical class label space* and are supposed to come from the *same domain*.
2. **Open-Set Semi-Supervised Learning** considers the sparsely labelled data and large sized auxiliary unlabelled data do not share the identical label space, i.e. there exists *class label mismatch* between the labelled and unlabelled datasets.
3. **Cross-Domain Semi-Supervised Learning** considers the labelled data and unlabelled data are drawn from different domains, where the two sets can exhibit both *data distributional drift* and *class label mismatch*.
4. **Unsupervised Learning** aims to learn purely from unlabelled data, without exploiting any task-related prior knowledge of label annotations for model training.

The ultimate objective of these learning paradigms is to advance the automatic exploitation of large-scale data with minimal human supervision. As summarised in Table 2.1, although these paradigms differ in how the training data is provided, they share a common principle in that auxiliary unsupervised learning objectives are generally formulated to learn from unlabelled training

Table 2.1: Comparison of different learning paradigms in the limited-label regime. SSL: semi-supervised learning; "✓": exist; "×": not exist; "–": may or may not exist.

| Type of Paradigm | Labelled Data | Class Label Mismatch | Domain Drift |
|---|---|---|---|
| Close-Set SSL | ✓ | × | × |
| Open-Set SSL | ✓ | ✓ | – |
| Cross-Domain SSL | ✓ | – | ✓ |
| Unsupervised Learning | × | × | × |

data. This is based on the premise that certain intrinsic structure presented in the unlabelled training set contains the useful information to be further utilised for inferring the underlying unknown class labels.

In this thesis, we study these *semi-supervised* and *unsupervised* learning paradigms in visual recognition, particularly in the tasks of visual classification and visual search. In the following, we review related work relevant to these paradigms in computer vision and machine learning, and discuss how our works differ from the previous state-of-the-art.

## 2.1 Close-Set Semi-Supervised Learning

**Semi-Supervised Learning** (SSL) [Zhu, 2005; Chapelle et al., 2009] has been studied for decades in the machine learning community, thanks to its promising impacts to effectively exploit large-scale unlabelled data together with sparsely labelled data. SSL is widely explored in various real-world application domains, such as image search [Fergus et al., 2009], medical data analysis [Papernot et al., 2017], web-page classification [Blum and Mitchell, 1998], document retrieval [Nigam and Ghani, 2000], genetics and genomics [Shi and Zhang, 2011; Libbrecht and Noble, 2015]. Existing works in SSL generally impose a *close-set* assumption in label space, i.e. all training samples – either labelled or unlabelled – belong to a set of pre-defined class labels. The generic idea of most existing works in SSL is to assign each unlabelled sample to a class label based on certain underlying data structure, e.g. manifold structure [Zhou et al., 2004; Weston et al., 2008], and graph structure [Zhu and Ghahramani, 2002]. Below, we review the most representative deep learning techniques that address close-set SSL for visual classification, including *consistency regularisation*, *entropy minimisation* and *adversarial training*.

***Consistency regularisation*** is a class of regularisation techniques that enforce the model out-

put consistency under variations in *input* or *weight* space. To enforce distributional smoothness under perturbation in *input space*, stochastic data augmentation [Sajjadi et al., 2016; Laine and Aila, 2017; Berthelot et al., 2019] or adversarial perturbation [Miyato et al., 2018] can be applied on the input to generate different transformations. For instance, Π-model [Sajjadi et al., 2016; Laine and Aila, 2017] encourages invariance in network outputs under random data augmentation. In virtual adversarial training (VAT) [Miyato et al., 2016, 2018], smoothness constraint is imposed to enforce the same prediction under virtual adversarial perturbation. To impose predictive consistency under variations in *weight space*, stochastic perturbation [Rasmus et al., 2015] or ensembling [Tarvainen and Valpola, 2017; Izmailov et al., 2018] can be adopted to generate outputs from non-identical models. For instance, Ladder Network [Rasmus et al., 2015] minimises a denoising cost derived between a clean forward pass and a noise corrupted forward pass from the network. Mean-Teacher [Tarvainen and Valpola, 2017] and Stochastic Weight Averaging (SWA) [Athiwaratkun et al., 2019] both apply ensembling through an exponential moving average (EMA) or equal average strategy in weight space to provide a more stable target for deriving a consistency cost. To summarise, an auxiliary consistency regularisation term is generally introduced to quantify the discrepancy between two predictive probability distributions or network outputs, which can be measured by Kullback-Leibler (KL) divergence or L1/L2-distance. By minimising such consistency regularisation loss term, the model is enforced to be invariant towards various data augmentation, perturbation, noisy corruption or ensembling.

***Entropy minimisation*** is a regularisation derived based on the *low density separation* principle [Grandvalet and Bengio, 2005; Chapelle et al., 2005], i.e. class decision boundary should be placed in the low density regions. This is accordant with the *cluster assumption* or *manifold assumption* in semi-supervised learning [Zhou et al., 2004; Weston et al., 2008], which hypotheses data points from the same class are likely to share the same *cluster* or *manifold* that is defined by measuring distances and densities. For deep model optimisation, entropy minimisation can simply be formulated as a regularisation penalty that minimises the entropy of model prediction [Lee, 2013; Miyato et al., 2018], which takes effect by enforcing a predictive distribution with low entropy (i.e. high confidence) to gradually assign each unlabelled sample to a certain class with highest probability.

***Adversarial training*** can generate *adversarial samples* along with auxiliary unsupervised learning signals for semi-supervised deep learning. According to whether the data density is ex-

plicitly modelled, this line of works can be categorised into two groups: *discriminative* and *generative* approaches. In *discriminative* adversarial training, as represented by virtual adversarial training (VAT) [Miyato et al., 2016, 2018], the data density is not explicitly modelled. To generate adversarial samples, small adversarial perturbations can be added to the inputs, while auxiliary supervision is imposed to enforce the output consistency under varying adversarial perturbations. In *generative* adversarial training, multiple variants of Generative Adversarial Network (GAN) [Springenberg, 2016; Salimans et al., 2016; Dumoulin et al., 2017; Dai et al., 2017] are proposed to leverage the joint effort of *unsupervised generative modelling* and *supervised discriminative training*. The generic idea is to utilise fake generated samples from the image generator together with real samples for semi-supervised learning. For instance, categorical GAN (CatGAN) [Springenberg, 2016] enforces a *certain distribution* (i.e. confident prediction with low entropy) on the real samples and an *uncertain distribution* (i.e. *uniform distribution* with high entropy) on the fake generated samples; while feature matching GAN [Salimans et al., 2016] labels the generated samples as an additional fake class ($K+1$) and the real samples as one of the real classes ($0 \sim K$) to train the discriminator classifier.

**Discussion.** As summarised in Figure 2.1, existing close-set SSL techniques can roughly be categorised into three groups. The key principle shared by these techniques is to propagate the class label assignments on unlabelled data for model optimisation, e.g. assigning soft targets on the unlabelled data or imposing uniform distributions on the synthetic data generated by GANs.

In Chapter 3, we propose a new SSL approach that is built upon a lightweight memory mechanism to propagate the class label assignments reliably based on the underlying manifold structure. Compared to existing SSL techniques, our approach is particularly efficient with **(1)** low computation cost, in contrast to adversarial training that requires to generate additional training data; and **(2)** low memory footprint, in contrast to most consistency regularisation techniques that either record a great amount of model predictions or need multiple forward network passes per iteration during training.

## 2.2 Open-Set Semi-Supervised Learning

**Open-Set Semi-Supervised Learning** is first identified as a challenging and understudied research problem in a realistic evaluation of semi-supervised learning (SSL) algorithms [Oliver et al., 2018], where existing SSL algorithms are found to degrade significantly when the unla-

belled training set contains out-of-distribution data, i.e. samples not present as known categories during training. For the first time, we systematically address this problem by a novel model formulation that can safely exploit the mixture of in- and out-of-distribution unlabelled training data [Chen et al., 2020b], which is detailed in Chapter 4. In the following, we review two research mainstreams that are closely related to open-set SSL, including *open-set recognition* and *out-of-distribution detection*.

### 2.2.1 Open-Set Recognition

**Open-set recognition**, also known as **open-world recognition** [Boult et al., 2019], aims to detect outliers (i.e. unknown categories) while classifying inliers (i.e. known categories) correctly at test time. This is a more realistic testing scenario compared to close-set recognition, as test samples are likely associated with noisy labels not present during training [Scheirer et al., 2012, 2014; Júnior et al., 2017; Bendale and Boult, 2016; Ge et al., 2018]. Along this branch of research, *OpenMax* [Bendale and Boult, 2016] and *G-OpenMax* [Ge et al., 2018] are two representative deep learning approaches formulated to reject samples of unknown categories. *OpenMax* is built upon deep networks and Extreme Value Theory, which fits a Weibull distribution over the activation values (computed from the penultimate layer of deep networks) of the positive training data per class, and estimates the Weibull CDF probability for each test sample to reject the unknown one. *G-OpenMax* (Generative-OpenMax) is further built upon *OpenMax*, which however, additionally introduces a conditional GAN that learns to generate samples conditioned on an additional label of unknown category ($K+1$), and exploits the generated images as auxiliary data apart from the data of known categories ($0 \sim K$) to train the network classifier.

### 2.2.2 Out-of-Distribution Detection

**Out-of-distribution (OOD) detection** targets at detecting samples lying out-of-distribution during test time, including *samples of unknown categories* and *adversarial samples* [Szegedy et al., 2014]. Unlike open-set recognition that requires to classify both known and unknown categories (i.e. *multi-class classification*), OOD detection mainly needs to distinguish whether the samples lie in- or out-of-distribution (i.e. *binary classification*). A naïve deep learning solution is to retrieve the maximum class probability from a softmax distribution [Hendrycks and Gimpel, 2017] as an indication of confidence (uncertainty) to detect OOD samples, which however, is likely to fail due to the overconfident tendency of predictions by deep networks [Nguyen et al.,

2015].

***Confidence calibration***, accordingly, addresses overconfidence by calibrating in the input, output, or weight space. As for processing in *input space* to mitigate overconfidence, *small perturbation* can be added at test time by back-propagating the gradient of the cross-entropy loss [Liang et al., 2018]. In *output space*, *temperature scaling* smooths the predictive distribution and therefore ameliorates the confidence score [Liang et al., 2018]. In addition, to avoid inference with overconfident targets, the output can be amended by regressing to *multiple dense representations* (i.e. word embeddings) of semantic labels [Shalev et al., 2018]. To constrain the mapping from *input to output space*, a *uniform distribution* can be imposed on the out-of-distribution data during training by exploiting auxiliary realistic data as *outlier exposure* [Hendrycks et al., 2019] or using GAN to generate synthetic data in low-density region [Lee et al., 2018a]. In the *weight space*, Bayesian probabilistic approaches – such as *Monte Carlo (MC) dropout* [Gal and Ghahramani, 2016], and *Stochastic Weight Averaging - Gaussian (SWAG)* [Maddox et al., 2019], can sample different weights following certain Gaussian distributions to perform Bayesian model averaging that basically results in more reliable predictive uncertainty. Another way is to *ensemble* multiple deep networks by averaging predictions [Lakshminarayanan et al., 2017] or majority vote [Shalev et al., 2018] to reduce bias and avoid overfitting the training data, which finally results in less overconfident model prediction.

**Discussion.** As summarised in Figure 2.2, *open-set recognition* and *out-of-distribution detection* (OOD) are two similar tasks aiming at detecting *unknown* samples at test time – although they are categorised as two different classification tasks: *multi-class* classification and *binary* classification. Compared to OOD, *Open-set recognition* is more challenging in the sense that more categories, including known and unknown ones, should all be correctly classified.

In Chapter 4, we consider a new learning paradigm, i.e. open-set semi-supervised learning (SSL). Unlike the aforementioned two tasks that focus on recognising unknown samples merely during *testing*, our goal is to enable semi-supervised *training* under the scenario where the labelled and unlabelled datasets do not share an identical class label space. This not only requires to detect unknown samples during training, but also urges for exploiting the unlabelled data in a reliable manner. Hence, open-set SSL is much more challenging than the standard close-set SSL due to the more sophisticated underlying class distribution of unlabelled data.

## 2.3 Cross-Domain Semi-Supervised Learning

**Cross-Domain Semi-Supervised Learning** – commonly known as ***domain adaptation*** – aims to transfer the prior knowledge learnt from the *labelled source domain* to the *unlabelled target domain*, as well as to tackle the *domain drift* between data collected from different domains. This is an essential problem in visual recognition, given that the statistical properties of visual data are quite sensitive to a wider variety of factors, e.g. illumination, viewpoint, resolution, occlusion, and background clutter induced by different scenes, camera characteristics, times of the day, and weather conditions. According to the availability of label annotations in target domain, domain adaptation strategies can be classified into *semi-supervised domain adaptation* and *unsupervised domain adaptation* – the former assumes the target data is partially labelled, while the latter considers the target data is completely unlabelled. Based on whether the label space is shared between domains, domain adaptation paradigms can also be grouped into *close-set domain adaptation* and *open-set domain adaptation* – the former assumes the source and target data share the same label space, while the latter considers the label space is not identical across two domains. Below, we review the two more challenging branches of research in domain adaptation, i.e. *unsupervised domain adaptation* and *open-set domain adaptation*.

### 2.3.1 Unsupervised Domain Adaptation

**Unsupervised domain adaptation** has been intensively studied due to its promising effect in avoiding the need of labelling target domain data. The generic aim is to learn domain-invariant feature representations that are immune to the domain discrepancy. To achieve this aim, a wide variety of deep learning techniques have been explored, as represented by *discriminative adversarial training*, *similarity metrics matching*, and *image style transfer*, which we describe next.

***Discriminative adversarial training*** is one of the most common and effective techniques for cross-domain distribution matching in latent representation space, which has been shown effective in domain adaptation [Ganin et al., 2016; Ganin and Lempitsky, 2015; Tzeng et al., 2015, 2017; Volpi et al., 2018; Long et al., 2018; Saito et al., 2018; Zhang et al., 2018b]. In a typical discriminative adversarial learning framework, a *domain discriminator* is generally introduced as a binary classifier to distinguish the domain labels of the features as either *source* or *target*; while a *feature generator network* is trained to fool the domain discriminator in an adversarial manner – In Domain Adversarial Neural Network (DANN) [Ganin et al., 2016], a *gradient reversal layer*

is introduced upon the feature generator to flip the gradients from the domain discriminator. To further advance this line of research, variants of improved feature generators and discriminators are proposed to improve discriminative feature matching. For instance, in Adversarial Discriminative Domain Adaptation (ADDA) [Tzeng et al., 2017], a target feature generate network is trained to mimic the distribution of the source feature generator network via the minimax game. In domain-invariant feature augmentation (DIFA) [Volpi et al., 2018], a target feature generator is trained adversarially to resemble the source features randomly generated from a pre-trained source feature generator. Besides improving the feature generators, domain discriminators can also be improved by introducing discriminative information into a class *conditional domain discriminator* (CDAN) [Long et al., 2018], and using *task-specific classifiers* as the domain discriminator (MCD) [Saito et al., 2018] or *multiple domain discriminators* at multi-level inside the network (CAN) [Zhang et al., 2018b].

***Similarity metrics matching*** imposes the feature distribution alignment constraints between the source and target domains by minimising the distributional divergence in the high-dimensional feature space [Tzeng et al., 2014; Long et al., 2015; Sun and Saenko, 2016; Koniusz et al., 2017; Xie et al., 2018]. This is achieved by introducing regularisation terms that measure the cross-domain distributional similarity, typically characterised by **(1)** *statistical means* – such as maximum mean discrepancy (MMD) [Long et al., 2015; Tzeng et al., 2014], and centroid alignment (MSTN) [Xie et al., 2018]; **(2)** *second- or higher-order statistics* – such as correlation alignment (CORAL) [Sun and Saenko, 2016], and within-class scatter alignment (So-HoT) [Koniusz et al., 2017]. By minimising certain similarity measure of the cross-domain divergence either in a per-batch [Long et al., 2015; Sun and Saenko, 2016] or per-class [Xie et al., 2018; Koniusz et al., 2017] manner, the feature encoder is trained to be domain-invariant.

***Image style transfer*** aims to mitigate the domain discrepancy at the pixel-level in a visually interpretable fashion, particularly handling the low-level domain drift caused by cross-domain variations in noise, resolution, illumination, and colour. In general, a cross-domain generative function is learnt to transform the source images into the target domain style [Yoo et al., 2016; Bousmalis et al., 2017; Shrivastava et al., 2017; Murez et al., 2018; Taigman et al.; Hoffman et al., 2018; Sankaranarayanan et al., 2018; Chen et al., 2019b]. Pixel-level adaptation could be more desirable than feature-level adaptation for multiple merits: **(1)** It provides better visible interpretability; **(2)** It maintains better pixel-level appearance details that are crucial for some

target tasks such as semantic segmentation [Hoffman et al., 2018; Sankaranarayanan et al., 2018; Chen et al., 2019b]; **(3)** It can generate abundant synthetic training data without labelling effort; and **(4)** It does not impose an unrealistic prior assumption that the target data have to share an identical label space as the source data. Given all these merits, different GAN models have been shown effective to learn a *real-to-real* [Yoo et al., 2016; Hoffman et al., 2018; Chen et al., 2019b] or *synthetic-to-real* [Shrivastava et al., 2017; Bousmalis et al., 2017; Sankaranarayanan et al., 2018] *mapping* for transforming the image style from one (i.e. source) to another (i.e. target), with learning objectives imposed to constrain the realism and semantics of the generated outputs. In particular, to constrain the image style, *adversarial loss* is commonly imposed to match the generated images with the target domain images. To preserve the semantic content, a wide variety of objectives can be introduced, such as *content loss* imposed at the feature space by an ImageNet pre-trained model [Gatys et al., 2016], *task-specific loss* imposed by the source task model [Hoffman et al., 2018; Chen et al., 2019b], and pixel-level *reconstruction loss* [Chen et al., 2019b] or *cycle consistency loss* [Hoffman et al., 2018] imposed to regularise the network training.

### 2.3.2 Open-Set Domain Adaptation

**Open-set domain adaptation** has attracted increasing research attention very recently [Busto and Gall; Saito et al.; Baktashmotlagh et al.; Liu et al., 2019]. Unlike most existing works in domain adaptation, this line of research considers the target domain contains samples of *unknown* categories, as opposed to close-set domain adaptation that assumes all target domain samples belong to one of the *known* categories in the source domain. To tackle this challenging task, two primary objectives are required be fulfilled: **(1)** detecting the *unknown* samples, while **(2)** improving recognition of the *known* categories in the target domain.

***Open-set adversarial training*** has been adopted as as an effective deep learning scheme for open-set domain adaptation [Saito et al.; Liu et al., 2019]. For instance, Adaptation by Backpropagation considers the unknown classes as an additional category $(K+1)$ in the classifier, which is trained to reject *unknown* target samples and align the *known* target samples with source samples through a *threshold-driven* adversarial learning scheme [Saito et al.]. Separate to Adapt introduces a set of per-class binary classifiers that learn to reject *unknown* target samples, while aligning the target and source data distributions in a weighted adversarial manner that discounts the degree of alignment according to the probabilities of *unknown* [Liu et al., 2019].

**Discussion.** As summarised in Figure 2.3, a wide variety of deep learning techniques have been explored to mitigate the domain drift across two domains at either the *feature-level* or *pixel-level*. While *feature-level* adaptation is generally more efficient to train with less computation cost when compared to *pixel-level* adaptation, the latter not only provides better visible interpretability with potentially richer visual data variations, but also does not impose any restrictive assumption on the target label space, thus facilitating more generic applicability to tackle open-set domain adaptation or pixel-level domain adaptation tasks such as semantic segmentation.

In Chapter 5, we address a more challenging open-set domain adaptation task, which consider the label spaces of the labelled source domain and unlabelled target domain are *completely disjoint*. This differs from existing works in open-set domain adaptation, which assume the label spaces of two domains are *partially overlapped*. To mitigate the domain gap in disjoint open-set domain adaptation, we propose a new instance-guided context rendering scheme to transfer the source domain label information into different target domain contexts using a novel dual conditional GAN. Our approach permits to produce abundant synthetic training data with rich contextual variations for domain-invariant representation learning.

## 2.4 Unsupervised Learning

**Unsupervised Learning** in computer vision considers to learning visual representation without exploiting any human-annotated labels. The pre-trained representation can be further transferred to unseen *downstream tasks* such as visual classification, detection, segmentation [He et al., 2020], or be directed deployed for visual tasks such as visual search [Chen et al., 2018a]. The most generic principle shared by existing work is to construct reliable surrogate supervision signals that avoid using any manual annotation of visual data, which can be formulated through *pretext tasks*, *discriminative clustering*, and *generative models*.

***Pretext tasks*** refer to tasks not directly related to the core learning task at hand, which therefore does not require any manual annotation for model training. By designing certain self-supervised learning objectives at either *pixel-level* or *image-level*, discriminative visual representation can be automatically learnt in a fully unsupervised manner. To formulate self-supervision at pixel level, *pixel-level pretext tasks* can be introduced to hallucinate colour values, denoise partial destructed noises, and inpaint missing region *per pixel* by L1/2 loss or adversarial loss, which are known as *colorization* [Zhang et al., 2017], *denoising* [Vincent et al., 2008], and *inpainting*

[Pathak et al., 2016] respectively. To impose self-supervision at image level, *image-level pretext tasks* generally targets at forming *pseudo labels* automatically *per image*. For instance, certain transformations can performed on the input image to **(1)** enforce invariance towards *examplar transformation per image* such as translation, rotation, scaling [Dosovitskiy et al.]; **(2)** classify the image rotations *per image* [Gidaris et al.]; **(3)** classify the patch orderings *per image patch* [Doersch et al.]; or **(4)** construct the training triplet loss based on temporal coherence guaranteed with *tracking* algorithm to discover positive and negative pairs [Wang and Gupta]. Although these self-supervised learning objectives generally does not align consistently with the target task learning objective, they permits to learn from contextual visual information that can implicitly enhance both visual discrimination and invariance, thus providing good model initialisation for the unseen downstream tasks.

***Discriminative clustering*** has long been used as effective unsupervised machine learning technique, as most represented by *k-means clustering* [Coates and Ng, 2012]. The generic idea of discriminative clustering is to divide data samples into a set of groups so as to reflect the underlying similarity and dissimilarity at the *group-level*, which can be adopted under deep learning frameworks (e.g. *DeepCluster* [Caron et al., a], *DeeperCluster* [Caron et al., b]) by iteratively grouping visual features using k-means and using the cluster assignments as pseudo class labels to train the networks. If one considers each instance as a class (cluster), self-supervision can be imposed per instance to encourage discrimination among individual instances. For example, one way of *instance-level* clustering is known as *invariant information clustering* [Ji et al., 2019], which maximises the mutual information between predictions of the original instance and the randomly perturbed instance obtained from data augmentation. *Instance discrimination* [Wu et al., 2018] is another way that employs a *memory bank* to record features of all the training samples, and uses noise-contrastive estimation [Gutmann and Hyvärinen, 2010] as the learning objective to ensure each sample is assigned to its own representation in the memory bank with a higher likelihood. To improve scalability to large dataset of billion-scale, *momentum contrast* [He et al., 2020] replaces *memory bank* in *instance discrimination* with a *dynamic dictionary* to enqueue the recent and dequeue the old mini-batches of samples, which finally permits contrastive learning by assigning each sample to the most similar representation in the *dynamic dictionary*.

***Generative models*** [Springenberg, 2016; Donahue et al.; Donahue and Simonyan] offer a new means of unsupervised learning by explicitly modelling the data distribution, as most represented

by Generative Adversarial Networks (GANs) [Goodfellow et al.]. Unsupervised visual features may come at the *discriminator-level* or *generator-level*. This is because, GANs not only contains a *discriminator* that could provide discriminative visual features to facilitate visual classification; but also offers an image *generator* that may serve as a powerful feature encoder to capture the semantics in its latent space. For instance, deep convolutional generative adversarial networks (*DCGAN*) [Radford et al., 2015] exploit the visual features from the convolutional discriminator to train additional SVM classifiers for visual classification. To exploit features from the generator, Bidirectional Generative Adversarial Networks (*BiGAN*) [Donahue et al.] is trained with a *joint discriminator loss* to tie the data and latent distribution together, which allows the image generators to capture semantic variation and offers useful feature representation for classification with simple One Nearest Neighbours (1NN). To further improve BiGAN, *BigBiGAN* [Donahue and Simonyan] adopts more powerful discriminator and generator architectures from BigGAN [Brock et al.] – together with additional *unary discriminator loss* to constrain the data or latent distribution independently, its unsupervised representation learning capability is boosted significantly.

**Discussion.** As summarised in Figure 2.4, existing unsupervised deep learning techniques can be roughly categorised into three groups. In general, unsupervised penalty can be formulated at different levels for model optimisation, which consequently can serve as the initialisation for supervised fine-tuning in various downstream tasks. This suggests an unsupervised pre-training objective does not necessarily need to be in line with the supervised fine-tuning objective. Unsupervised pre-training is quite attractive in the sense that generic supervised pre-training on large-scale labelled dataset such as ImageNet could be potentially avoided.

In Chapter 6, we study a special unsupervised learning scenario where the target task objective *is in line with* the unsupervised learning objective. This is typically useful in visual search tasks, such as person re-identification, where the training and test objectives both aim to match similar items of the same category based on certain metrics, such as L1 or L2 distance. To tackle this special unsupervised learning task, we proposed a novel deep association learning scheme that learns without any label annotations by associating nearest neighbour instances in the feature space through two types of ranking consistency.

## 2.5 Summary

The preceding sections have reviewed and discussed relevant and representative studies in the machine learning and computer vision literature, which collectively serve as the foundational knowledge for this thesis. Built upon these related works, this thesis presents four different deep learning paradigms in the limited-label regime, which however, all differ from the aforementioned studies in terms of problem formulation and (or) model formulation, as outlined below.

1. [Chapter 3] **Semi-supervised deep learning with memory**: Semi-supervised deep learning aims to learn jointly from sparsely labelled data and abundant unlabelled data by a joint optimisation of supervised and unsupervised learning objectives (Eq. 2.1).

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \alpha \mathcal{L}_{\text{unsupervised}} \tag{2.1}$$

   where $\alpha$ is a hyper-parameter, i.e. a scalar value tuned on the validation set. We introduce a new formulation of unsupervised learning objective based on a memory module, which propagates the soft assignments on all the data to progressively fit the labelled and unlabelled data towards the underlying manifold.

2. [Chapter 4] **Open-set semi-supervised learning by uncertainty-aware self-distillation**: Open-set semi-supervised learning considers semi-supervised learning under class distribution mismatch, i.e. the labelled data and unlabelled data do not share the same class label space. This is a new semi-supervised learning problem that has not be effectively addressed before. To optimise the model without introducing corrupted learning signals, it is essential to discard the unlabelled data that lie out of distribution by imposing the unsupervised learning objective selectively (Eq. 2.2).

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \alpha(\cdot) \mathcal{L}_{\text{unsupervised}} \tag{2.2}$$

   where $\alpha(\cdot)$ is a weighting function defined to discount the importance of the unsupervised learning objective when an unlabelled sample is likely lying out-of-distribution. We propose a new model formulation: uncertainty-aware self-distillation. Our model ensembles all the historic network forward propagated predictions on the fly to derive soft label assignments on all the training data, which yields a reliable confidence estimate that could discard out-of-distribution samples to minimise their negative impact in optimisation.

3. [Chapter 5] **Open-set cross-domain learning by instance-guided context rendering**: Open-set cross-domain semi-supervised learning targets at mitigating the cross-domain discrepancy between the labelled source data and unlabelled target data, where the two sets of data do not share the same class space. A domain adaptation model is typically optimised by a supervised learning objective imposed on the source domain data and an unsupervised learning objective to minimise the cross-domain discrepancy (Eq. 2.3).

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \alpha \mathcal{L}_{\text{cross-domain}} \tag{2.3}$$

where $\alpha$ is a hyper-parameter. We introduce a new pixel-level domain adaptation model: instance-guided context rendering, which transfers the source domain labels into the target domain contexts by image style transfer. Our model produces abundant synthetic training data to learn an more expressive representation that is both discriminative to task-relevant information (i.e. class labels) and invariant to task-irrelevant factors (i.e. contexts).

4. [Chapter 6] **Unsupervised learning by online deep association**: Unsupervised learning seeks to learn meaningful representations without relying on any manual annotation. As label information is missing during training, a discriminative unsupervised model is typically optimised by preforming and (or) discovering certain pseudo labels that are not necessarily related to the learning task at hand (Eq. 2.4).

$$\mathcal{L} = \mathcal{L}_{\text{unsupervised}} \tag{2.4}$$

where $\mathcal{L}_{\text{unsupervised}}$ may involve multiple unsupervised loss terms formulated to achieve different objectives. We formulate a new unsupervised learning scheme: deep association learning, which employs two types of unsupervised consistency constraints to gradually associate image representation to video representation and associate the video representations across different views. Our model presents two ways of learning with pseudo labels either preformed based on certain information or discovered dynamically during training.

In summary, the essential key idea shared by all our works lies at formulating the proper unsupervised learning objectives that could learn from unlabelled visual data either in a semi-supervised or unsupervised manner, as summarised in Eq. 2.1, 2.2, 2.3, 2.4. Based on whether there is labelled data, or whether there is class label mismatch and domain drift between the labelled and unlabelled datasets (Table 2.1), we introduce different deep learning paradigms and

review their related works in Section 2.1, 2.2, 2.3, 2.4. We also briefly outline our problem and model formulations in Section 2.5, which are detailed in the following chapters.

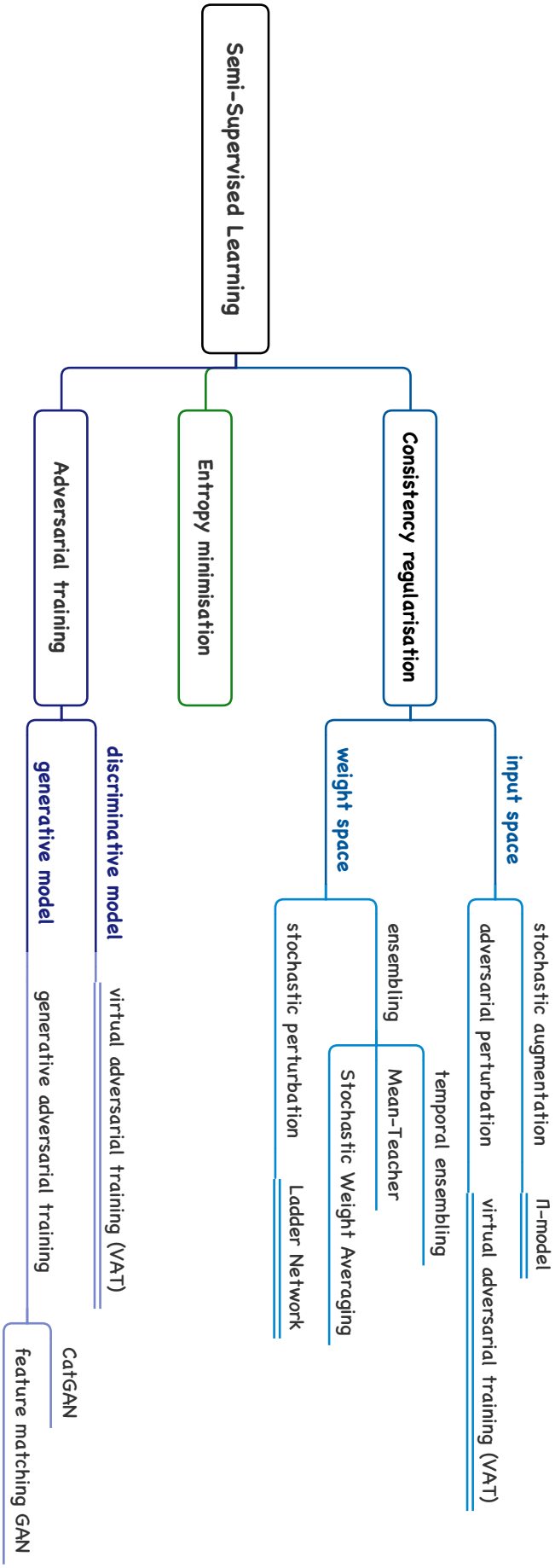



Figure 2.1: A taxonomic summary of representative semi-supervised deep learning techniques.

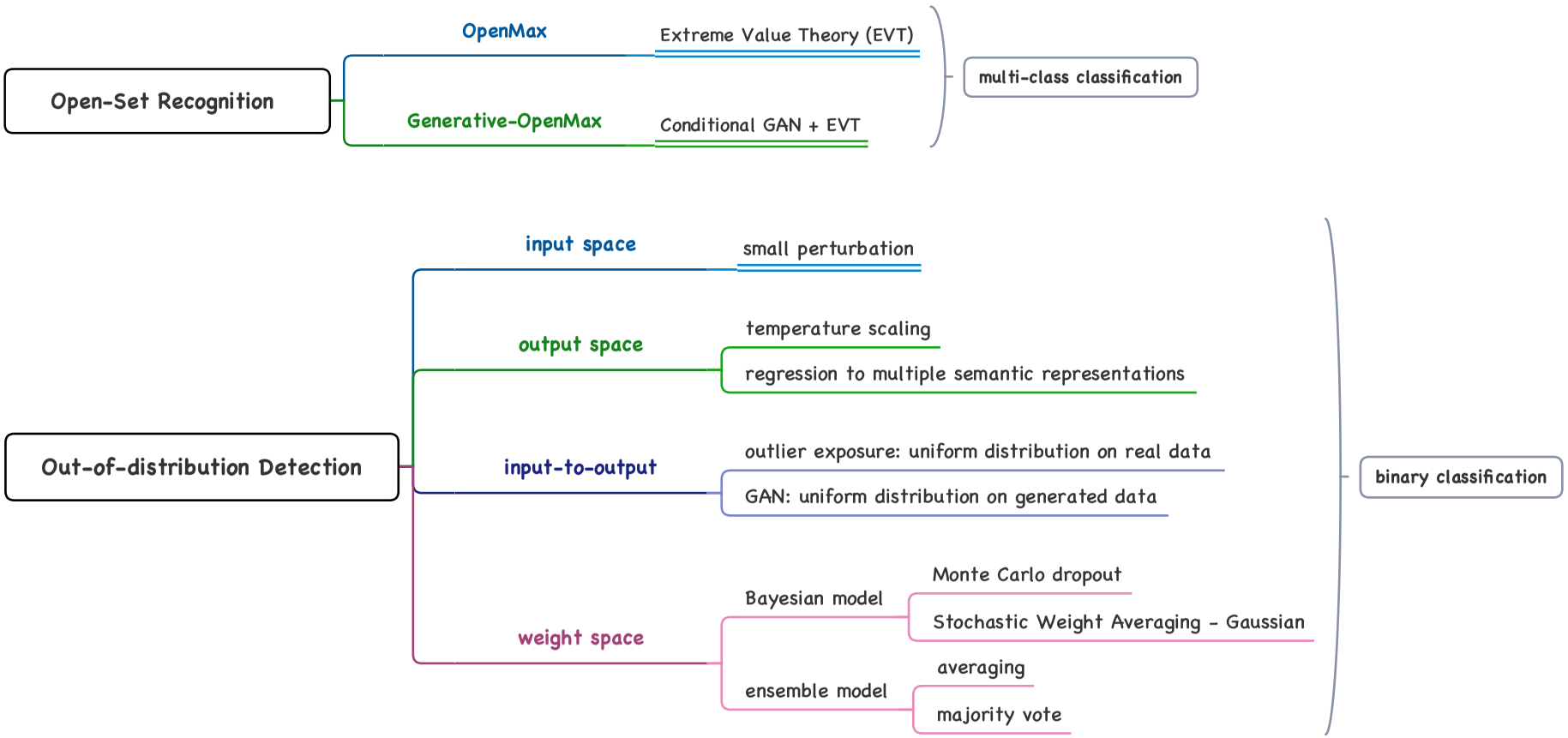



Figure 2.2: A taxonomic summary of representative deep learning techniques for open-set recognition and out-of-distribution (OOD) detection.

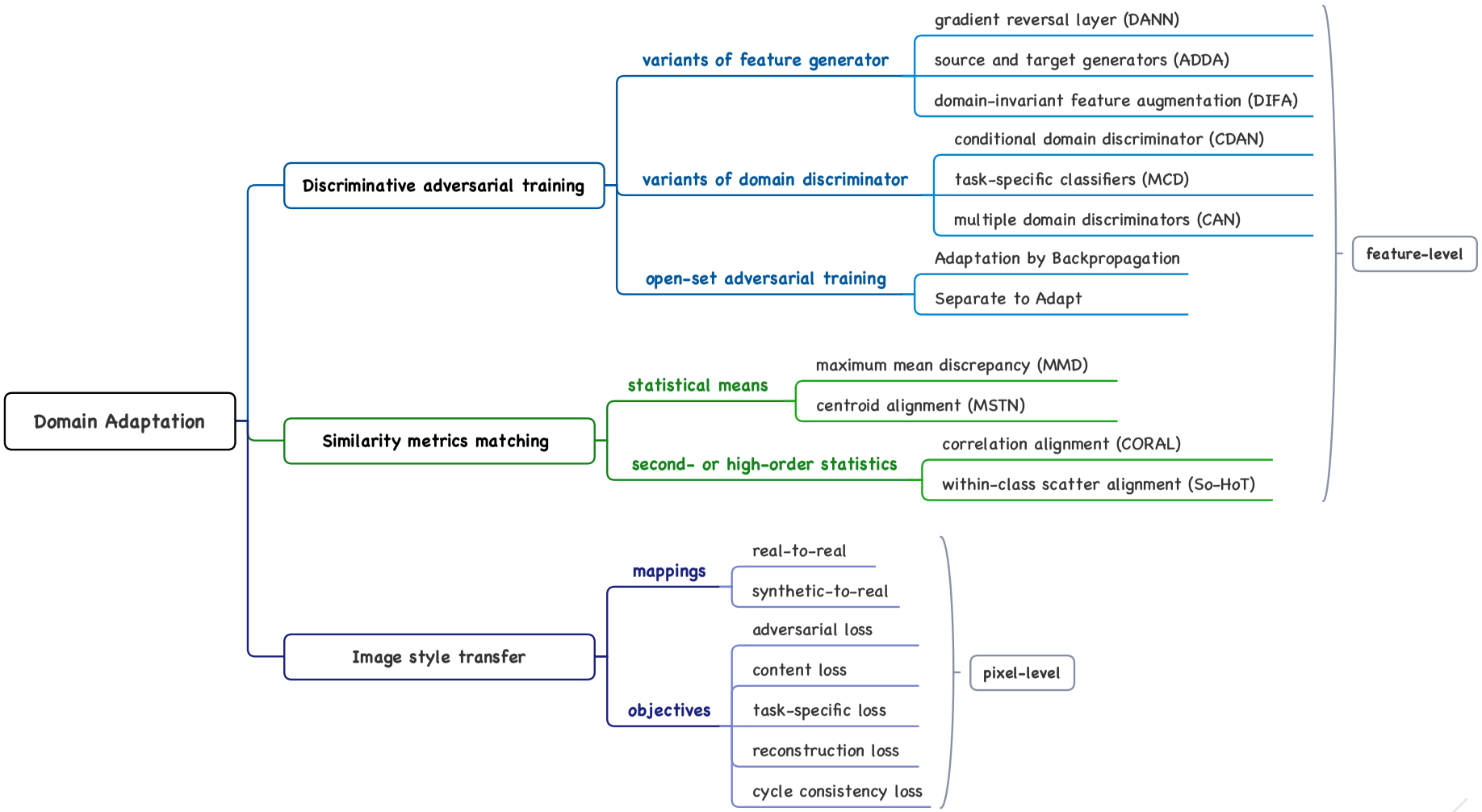



Figure 2.3: A taxonomic summary of representative cross-domain learning techniques, which mitigate the cross-domain discrepancy at either *feature-level* or *pixel-level*.

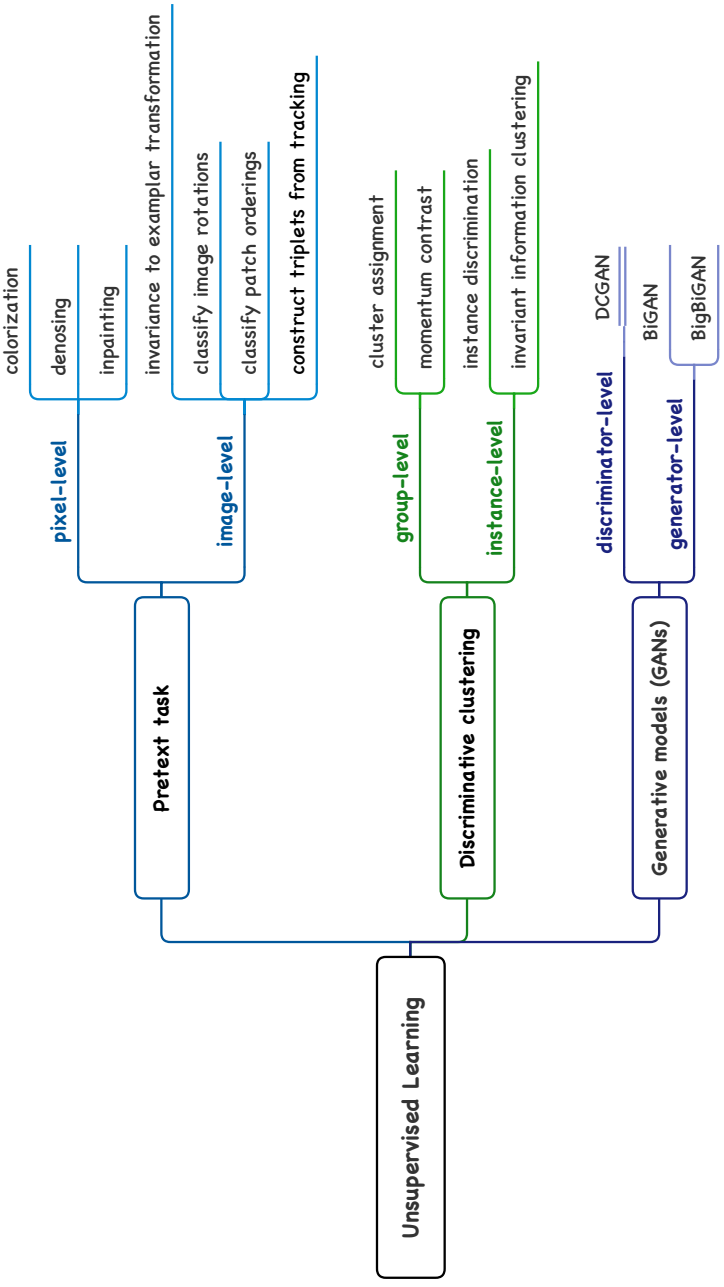



Figure 2.4: A taxonomic summary of representative unsupervised deep learning techniques.

# Chapter 3

# Semi-Supervised Deep Learning with Memory

---

Semi-supervised learning (SSL) aims to boost the model performance by utilising the large amount of unlabelled data when only a limited amount of labelled data is available [Chapelle et al., 2009; Zhu, 2005]. It is motivated that unlabelled data are available at large scale but labelled data are scarce due to high labelling costs. This learning scheme is useful and beneficial for many applications such as image search [Fergus et al., 2009], web-page classification [Blum and Mitchell, 1998], document retrieval [Nigam and Ghani, 2000], genomics [Shi and Zhang, 2011], and so forth. In the SSL literature, the most straightforward SSL algorithm is self-training where the target model is incrementally trained by additional self-labelled data given by the model's own predictions with high confidence [Nigam and Ghani, 2000; Blum and Mitchell, 1998; Rosenberg et al., 2005]. This method is prone to error propagation in model learning due to wrong predictions of high confidence. Other common methods include Transductive SVM [Joachims, 1999; Chapelle et al., 2005] and graph-based methods [Zhu et al., 2003; Blum et al., 2004], which, however, are likely to suffer from poor scalability to large-scale unlabelled data due to inefficient optimisation.

Recently, neural network based SSL methods [Ranzato and Szummer, 2008; Weston et al., 2008; Lee, 2013; Kingma et al., 2014; Springenberg, 2016; Rasmus et al., 2015; Miyato et al., 2016; Sajjadi et al., 2016; Maaløe et al., 2016; Haeusser et al., 2017; Tarvainen and Valpola, 2017] start to dominate the progress due to the powerful representation-learning ability of deep neural networks. A common strategy is to train the deep neural networks by simultaneously optimising a standard supervised classification loss on labelled samples along with an additional

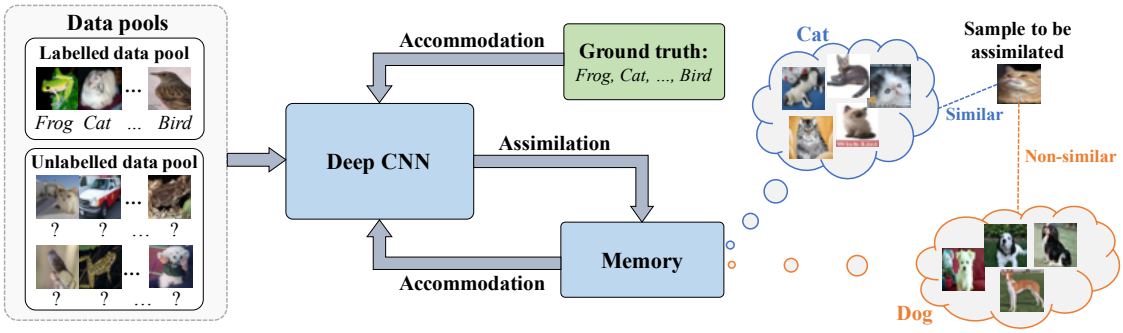



Figure 3.1: Illustration of the memory-assisted semi-supervised deep learning framework that integrates a deep CNN with an external memory module trained concurrently. The memory module assimilates the incoming training data on-the-fly and generates an additional unsupervised memory loss to guide the network learning along with the standard supervised classification loss.

unsupervised loss term imposed on either unlabelled data [Lee, 2013; Salimans et al., 2016; Dumoulin et al., 2017] or both labelled and unlabelled data [Weston et al., 2008; Rasmus et al., 2015; Miyato et al., 2016; Laine and Aila, 2017]. These additional loss terms are considered as *unsupervised* supervision signals since no ground-truth label is required for loss computation. For example, Lee [Lee, 2013] utilises the cross-entropy loss computed on the pseudo labels (the classes with the maximum predicted probability given by the up-to-date network) of unlabelled samples. Rasmus et al. [Rasmus et al., 2015] adopt the reconstruction loss between one clean forward propagation and one stochastically-corrupted forward propagation derived for the same sample. Miyato et al. [Miyato et al., 2016] define the distributional smoothness against local random perturbation as an unsupervised penalty. Laine et al. [Laine and Aila, 2017] introduce an unsupervised $L_2$ loss to penalise the inconsistency between the network predictions and the temporally ensembled network predictions. Tarvainen et al. [Tarvainen and Valpola, 2017] formulate an unsupervised $L_2$ loss to regress the network predictions to the model predictions derived from the Mean Teacher – an ensemble model that keeps an exponential moving average of model weights during training. Overall, the rationale of these SSL algorithms is to regularise the network by enforcing smooth and consistent classification boundaries that are robust to random perturbation [Rasmus et al., 2015; Miyato et al., 2016]; or to enrich the supervision signals by exploiting the knowledge learned by the network, such as using pseudo labels [Lee, 2013] and temporally ensembled predictions [Laine and Aila, 2017].

Most of the aforementioned methods typically utilise the up-to-date in-training network to formulate an additional unsupervised penalty that enables semi-supervised learning. We consider that this kind of deep SSL scheme is sub-optimal, given that the memorising capacity of

deep networks is often incomplete and insufficiently compartmentalised to represent knowledge accrued in the past learning iterations. To effectively leverage such knowledge, we introduce a memory mechanism into the deep network training process to enable semi-supervised learning from small-sized labelled and large-sized unlabelled training data. Our proposed memory mechanism is related to the class of learning model *memory networks* proposed by Weston et al. [Weston et al., 2014], which allows read and write operation to continually retrieve and update information in the memory. While the vanilla *memory networks* are formulated to store the linguistic knowledge and predict answers in a question answering task, our proposed memory mechanism, on the other hand, stores the continually refreshing class-level feature representations and network predictions to derive class assignment for the unlabelled samples. In spirit of the Piaget's theory on human's ability of *continual learning* [Ginsburg and Opper, 1988], we aim to design a SSL scheme that permits the deep model to additionally learn from its memory (*assimilation*) and adjust itself to fit optimally the incoming training data (*accommodation*) in an incremental manner. To this end, we formulate a novel memory-assisted semi-supervised deep learning framework: Memory-Assisted Deep Neural Network (MA-DNN) as illustrated in Figure 6.1. MA-DNN is characterised by an assimilation-accommodation interaction between the network and an external memory module.

Augmenting a network with an external memory component is attractive due to its flexible capability of storing, abstracting and organising the past knowledge into a structural and addressable form, which have been widely adopted to a variety of challenging tasks such as question answering [Weston et al., 2014; Sukhbaatar et al., 2015; Miller et al., 2016] and one-shot learning [Santoro et al., 2016; Kaiser et al., 2017]. As earlier works, Weston et al. [Weston et al., 2014] propose Memory Networks, which integrate inference components with a memory component that can be read and written to remember supporting facts from the past for question answering. Kaiser et al. [Kaiser et al., 2017] propose a life-long memory module to record network activations of rare events for one-shot learning. Our work is conceptually inspired by these works, but it is the **first attempt** to explore the **memory mechanism** in semi-supervised deep learning. Besides the basic storage functionality, our memory module induces an assimilation-accommodation interaction to exploit the memory of model learning and generate an informative unsupervised memory loss that permits semi-supervised learning.

Specifically, the key to our framework design is two-aspect: (1) the class-level discriminative

feature representation and the network inference uncertainty are gradually accumulated in an external memory module; (2) this memorised information is utilised to assimilate the newly incoming image samples on-the-fly and generate an informative unsupervised memory loss to guide the network learning jointly with the supervised classification loss.

In summary, our **contribution** in this work is two-fold:

- We propose to exploit the *memory* of model learning to enable semi-supervised deep learning from the sparse labelled and abundant unlabelled training data, whilst fully adopting the existing end-to-end training process. This is in contrast to most existing deep SSL methods that typically ignore the memory of model learning.

- We formulate a novel *Memory-Assisted Deep Neural Network* (MA-DNN) characterised by a memory mechanism. We introduce an unsupervised memory loss compatible with the standard supervised classification loss to enable semi-supervised learning. Extensive comparative experiments demonstrate the advantages of our proposed MA-DNN model over a wide variety of state-of-the-art semi-supervised deep learning methods.

## 3.1 Memory-Assisted Deep Neural Network

We consider semi-supervised deep learning in the context of multi-class image classification. In this context, we have access to a limited amount of labelled image samples $\mathcal{D}_L = \{(\boldsymbol{I}_{i,l}, \boldsymbol{y}_{i,l})\}_i^{n_l}$ but an abundant amount of unlabelled image samples $\mathcal{D}_U = \{(\boldsymbol{I}_{i,u})\}_i^{n_u}$, where $n_u \gg n_l$. Each unlabelled image is assumed to belong to one of the same $K$ object categories (classes) $\mathcal{Y} = \{\boldsymbol{y}_i\}_i^K$ as the labelled data, while their ground-truth labels are not available for training. The key objective of SSL is to enhance the model performance by learning from the labelled image data $\mathcal{D}_L$ and the additional unlabelled image data $\mathcal{D}_U$ simultaneously. To that end, we formulate a memory-assisted semi-supervised deep learning framework that integrates a deep neural network with a memory module, We call this *Memory-Assisted Deep Neural Network* (MA-DNN).

### 3.1.1 Approach Overview

The overall design of our MA-DNN architecture is depicted in Figure 3.2. The proposed MA-DNN contains three parts: **(1)** A deep neural network (Section 3.1.2); **(2)** A memory module designed to record the memory of model learning (Section 3.2); and **(3)** An assimilation-

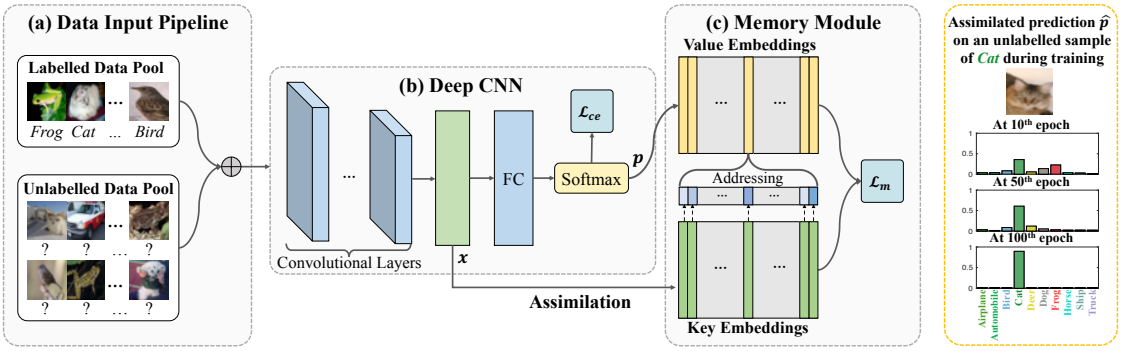



Figure 3.2: An overview of Memory-Assisted Deep Neural Network (MA-DNN) for semi-supervised deep learning. During training, given **(a)** sparse labelled and abundant unlabelled training data, mini-batches of labelled/unlabelled data are feed-forward into **(b)** the deep CNN to obtain the up-to-date feature representation $\boldsymbol{x}$ and probabilistic prediction $\boldsymbol{p}$ for each sample. Given **(c)** the updated memory module, memory assimilation induces another multi-class prediction $\hat{\boldsymbol{p}}$ (Eq. (3.4)) for each sample via key addressing and value reading. In accommodation, a memory loss $\mathcal{L}_m$ (Eq. (3.7)) is computed from $\hat{\boldsymbol{p}}$ and employed as an additional supervision signal to guide the network learning jointly with the supervised classification loss. At test time, the memory module is no longer needed, so it does not affect the deployment efficiency.

accommodation interaction mechanism introduced for effectively exploiting the memory to facilitate the network optimisation in semi-supervised learning (Section 3.2.1).

### 3.1.2 Network Architecture Selection

The proposed framework aims to work with existing standard deep neural networks. We select the Convolutional Neural Network (CNN) in this work due to its powerful representation-learning capability for imagery data. To train a CNN for image classification, the supervised cross-entropy loss function is usually adopted. During training, given any training sample $\boldsymbol{I}$, we feed-forward it through the up-to-date deep network to obtain a feature vector $\boldsymbol{x}$ and a multi-class probabilistic prediction vector $\boldsymbol{p}$ over all classes. Specifically, the $j$-th class posterior probability of the labelled image sample $\boldsymbol{I}_i$ as

$$p(\boldsymbol{y}_j|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{W}_j^\top \boldsymbol{x}_i)}{\sum_{j=1}^{|\mathcal{Y}|} \exp(\boldsymbol{W}_j^\top \boldsymbol{x}_i)} \tag{3.1}$$

where $\boldsymbol{x}_i$ refers to the embedded deep feature representation of $\boldsymbol{I}_i$ extracted by the deep CNN, and $\boldsymbol{W}_j$ is the $j$-th class prediction function parameter. The cross-entropy loss on $\boldsymbol{I}_i$ against the ground-truth class label $k$ is computed as

$$\mathcal{L}_{\text{ce}} = -\sum_{j=1}^{K} \mathbb{1}[\boldsymbol{y}_j = k] \log\Big(p(\boldsymbol{y}_j|\boldsymbol{x}_i)\Big) \tag{3.2}$$

Obviously, the cross-entropy loss function is restricted to learn from the labelled samples alone. To take advantage of the unlabelled training samples, a straightforward way is to utilise the

predicted labels given by the up-to-date model in training. This, however, may be error-prone and unreliable given immature label estimations particularly at the beginning of model training. This presents a catch-22 problem. We overcome this problem by introducing a memory module into the network training process to progressively estimate more reliable predictions on the unlabelled data.

## 3.2 Memory Module

To take advantage of the memorisable information generated in model learning, it is necessary for us to introduce a memory module. We consider two types of memory experienced by the network-in-training: **(1)** the class-level feature representation, and **(2)** the model inference uncertainty.

To manage these memorisable information, we construct the memory module in a key-value structure [Miller et al., 2016]. The memory module consists of multiple slots with each slot storing a symbolic pair of (*key*, *value*). In particular, the key embedding is the dynamically updated *feature representation* of each class in the feature space. Utilising an univocal representation per class is based on the assumption that deep feature embeddings of each class can be gradually learned to distribute around its cluster centroid in the feature space [Wen et al., 2016]. Based on this assumption, the global feature distribution of all classes is represented by their cluster centroids in the feature space, whilst these cluster centroids are cumulatively updated as the key embeddings in a batch-wise manner. On the other hand, the value embedding records the similarly updated *multi-class probabilistic prediction* w.r.t. each class. Hence, each value embedding is the accumulated network predictions of samples from the same class that encodes the overall model inference uncertainty at the class level.

To represent the incrementally evolving feature space and the up-to-date overall model inference uncertainty, memory update is performed every iteration to accomodate the most recent updates of the network. We only utilise the labelled data for memory update, provided that unlabelled samples have uncertainty in class assignment and hence potentially induce the risk of error propagation. Formally, suppose there exist $n_j$ labelled image samples $\{\boldsymbol{I}_i\}$ from the $j$-th class ($j \in \{1, \cdots, K\}$) with their feature vectors and probabilistic predictions as $\{\boldsymbol{x}_i, \boldsymbol{p}_i\}$, the $j$-th

memory slot $(\boldsymbol{k}_j, \boldsymbol{v}_j)$ is cumulatively updated over all the training iterations as follows.

$$\begin{cases} \boldsymbol{k}_j \leftarrow \boldsymbol{k}_j - \eta \nabla \boldsymbol{k}_j \\ \boldsymbol{v}_j \leftarrow \dfrac{\boldsymbol{v}_j - \eta \nabla \boldsymbol{v}_j}{\sum_{i=1}^{K} (\boldsymbol{v}_{j,i} - \eta \nabla \boldsymbol{v}_{j,i})} \end{cases} \text{with} \begin{cases} \nabla \boldsymbol{k}_j = \dfrac{\sum_{i=1}^{n_j} (\boldsymbol{k}_j - \boldsymbol{x}_i)}{1 + n_j} \\ \nabla \boldsymbol{v}_j = \dfrac{\sum_{i=1}^{n_j} (\boldsymbol{v}_j - \boldsymbol{p}_i)}{1 + n_j} \end{cases} \tag{3.3}$$

where $\eta$ denotes the learning rate (set to $\eta = 0.5$ in our experiments). The value embedding $\boldsymbol{v}_j$ is normalised to ensure its probability distribution nature. Along the training process, as the gradients $(\nabla \boldsymbol{k}_j, \nabla \boldsymbol{v}_j)$ progressively get smaller, the key and value embeddings will become more reliable to reflect the underlying feature structures and multi-class distributions. To begin the training process without imposing prior knowledge, we initialise all the key and value embeddings to $\mathbf{0}$ and $\frac{1}{K} \cdot \mathbf{1}$ (a uniform probabilistic distribution over $K$ classes), respectively. This indicates the memorised information is fully discovered by the network during training, without any specific assumption on the problem settings, therefore potentially applicable to different semi-supervised image classification tasks.

### 3.2.1 The Assimilation-Accomodation Interaction

Given the updated memory of model learning, we further employ it to enable semi-supervised deep learning. This is achieved by introducing an assimilation-accomodation interaction mechanism with two operations executed every training iteration: **(1)** *Memory Assimilation*: Compute the memory prediction for each training sample by key addressing and value reading; **(2)** *Accommodation*: Compute the memory loss to formulate the final semi-supervised learning objective. We present the details of these operations in the following.

**(1) Memory Assimilation.** Given the forward propagated image representation $\boldsymbol{x}$ and network prediction $\boldsymbol{p}$ of the image $\boldsymbol{I}$, memory assimilation induces another multi-class probabilistic prediction $\hat{\boldsymbol{p}}$ based on the updated memory. We obtain this by *key addressing* and *value reading* [Miller et al., 2016]. Specifically, key addressing is to compute the addressing probability $w(\boldsymbol{m}_i|\boldsymbol{I})$, i.e. the probabilistic assignment to each memory slot $\boldsymbol{m}_i = (\boldsymbol{k}_i, \boldsymbol{v}_i)$, $i \in \{1, \cdots, K\}$, based on pairwise similarity w.r.t. each key embedding. In essence, $w(\boldsymbol{m}_i|\boldsymbol{I})$ is the cluster assignment in the feature space. Given the addressing probabilities over all $K$ memory slots, value reading is then applied to compute the memory prediction $\hat{\boldsymbol{p}}$ by taking a weighted sum of all the value embeddings as follows.

$$\hat{\boldsymbol{p}} = \sum_{i=1}^{K} w(\boldsymbol{m}_i|\boldsymbol{I}) \, \boldsymbol{v}_i \tag{3.4}$$

According to label availability, we adopt two addressing strategies. The first is *position-based* addressing applied to labelled training samples. Formally, suppose the training sample $\boldsymbol{I}$ is labelled as the $k$-th class, the addressing probability is attained based on the position $k$ as

$$w(\boldsymbol{m}_i|\boldsymbol{I}) = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \tag{3.5}$$

The second is *content-based* addressing applied to unlabelled image samples. This strategy computes the addressing probability based on the pairwise similarity between the image sample $\boldsymbol{I}$ and the key embeddings $\boldsymbol{k}_i$ as

$$w(\boldsymbol{m}_i|\boldsymbol{I}) = \frac{e^{-\text{dist}(\boldsymbol{x},\boldsymbol{k}_i)}}{\sum_{j=1}^{K} e^{-\text{dist}(\boldsymbol{x},\boldsymbol{k}_j)}} \tag{3.6}$$

where $\boldsymbol{x}$ is the extracted feature vector of $\boldsymbol{I}$ and dist() denotes the Euclidean distance. Eq. (3.6) can be considered as a form of label propagation [Zhu and Ghahramani, 2002] based on the *cluster assumption* [Weston et al., 2008; Zhou et al., 2004], in the sense that the probability mass is distributed according to proximity to each cluster centroid in the feature space. That is, probabilistic assignments are computed based on cluster memberships.

**(2) Accommodation.** This operation provides the deep network with a memory loss to formulate the final semi-supervised learning objective such that the network can learn additionally from the unlabelled data. Specifically, we introduce a *memory loss* on each training sample $\boldsymbol{x}$ as follows.

$$\mathcal{L}_m = H(\hat{\boldsymbol{p}}) + \max(\hat{\boldsymbol{p}}) D_{\text{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}}) \tag{3.7}$$

where $H()$ refers to the entropy measure; max() is the maximum function that returns the maximal value of the input vector; $D_{KL}()$ is the Kullback-Leibler (KL) divergence. Both $H()$ and $D_{KL}()$ can be computed without ground-truth labels and thus applicable to semi-supervised learning. The two loss terms in Eq. (3.7) are named as the Model Entropy (ME) loss and the Memory-Network Divergence (MND) loss, as explained below.

**(i)** The Model Entropy (ME) loss term $H(\hat{\boldsymbol{p}})$ is formally computed as

$$H(\hat{\boldsymbol{p}}) = -\sum_{j=1}^{K} \hat{\boldsymbol{p}}(j) \log\ \hat{\boldsymbol{p}}(j) \tag{3.8}$$

which quantifies the amount of information encoded in $\hat{\boldsymbol{p}}$. From the information-theoretic perspective, the entropy reflects the overall model inference uncertainty. A high entropy on a *labelled* image sample indicates that $\hat{\boldsymbol{p}}$ is an ambiguous multimodal probability distribution, which

corresponds to the retrieved value embedding of a specific class. This indicates that the network cannot well distinguish between this class and the other classes, which is resulted from assigning inconsistent probabilistic predictions to image samples within the same class. On the other hand, a high entropy on an *unlabelled* sample suggests the severe class distribution overlap between different classes in the feature space. This is because the unlabelled sample cannot be assigned to a certain class with high probability. Therefore, minimising the model entropy $H$ is equivalent to reducing class distribution overlap in the feature space and penalising inconsistent network predictions at the class level, which is essentially motivated by the *entropy minimisation* principle [Grandvalet and Bengio, 2005]. It is worth mentioning that minimising $H(\hat{\boldsymbol{p}})$ differs from minimising $H(\boldsymbol{p})$. While the latter blindly enforces a confident model prediction on the samples, the former encourages the sample to assign to its most similar class in the feature space with higher similarity.

**(ii)** The Memory-Network Divergence (MND) loss term $D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}})$ is computed between the network prediction $\boldsymbol{p}$ and the memory prediction $\hat{\boldsymbol{p}}$ as follows.

$$D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}}) = \sum_{j=1}^{K} \boldsymbol{p}(j) \log \frac{\boldsymbol{p}(j)}{\hat{\boldsymbol{p}}(j)} \tag{3.9}$$

$D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}})$ is a non-negative penalty that measures the discrepancy between two distributions: $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$. It represents the additional information encoded in $\boldsymbol{p}$ compared to $\hat{\boldsymbol{p}}$ in information theory. Minimising this KL divergence prevents the network prediction from overly deviating from the probabilistic distribution derived from the memory module. When $D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}}) \rightarrow 0$, it indicates the network predictions match well with its memory predictions. Additionally, we also impose a dynamic weight: $\max(\hat{\boldsymbol{p}})$, the maximum probability value of $\hat{\boldsymbol{p}}$, to discount the importance of $D_{\mathrm{KL}}()$ when given an ambiguous memory prediction (multimodal probability distribution). Hence, $\boldsymbol{p}$ is encouraged to match with $\hat{\boldsymbol{p}}$ only when $\hat{\boldsymbol{p}}$ corresponds to a confident memory prediction (peaked probability distribution).

The final **semi-supervised learning objective function** is formulated by merging Eq. (3.7) and Eq. (3.2) as follows.

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{m} \tag{3.10}$$

where $\lambda$ is a hyper-parameter that is set to 1 to ensure equivalent importance of two loss terms during training.

*Remark.* The fundamental role of memory module is to record the class-level feature representations and model inference uncertainty, such that labels can be propagated from the known feature

distribution to the unlabelled samples. In essence, a Gaussian assumption is implicitly imposed, where the key embedding can be considered as the estimated mean of the Gaussian distribution while the value embedding reflect the uncertainty of the estimated mean. As the network weights are constantly updated, it is also essential to constantly update the estimated per-class mean and uncertainty to capture the updated underlying feature distributions.

### 3.2.2 Model Training

The proposed MA-DNN is trained by standard Stochastic Gradient Descent algorithm in an end-to-end manner. The algorithmic overview of model training is summarised in Algorithm 1.

---

**Algorithm 1** Memory-Assisted Semi-Supervised Deep Learning.

---

**Input:** Labelled data $\mathcal{D}_L$ and unlabelled data $\mathcal{D}_U$.

**Output:** A deep CNN model for classification.

**for** $t = 1$ **to** *max_iter* **do**

Sampling a mini-batch of labelled & unlabelled data.

Network forward propagation (samples feed-forward).

Memory update (Eq. (3.3)).

Network supervised loss computation (Eq. (3.2)).

Memory assimilation (Eq. (3.4)) and accommodation (Eq. (3.7)).

Network update by back-propagation (Eq. (6.7)).

**end for**

---

## 3.3 Discussion

Whilst sharing the generic spirit of introducing an unsupervised penalty as prior works in semi-supervised learning [Lee, 2013; Salimans et al., 2016; Dumoulin et al., 2017; Weston et al., 2008; Rasmus et al., 2015; Miyato et al., 2016; Laine and Aila, 2017], our method is **novel** in a number of fundamental ways: **(I)** Exploiting the memory of model learning: Instead of relying on the incomplete knowledge of a single up-to-date network to derive the additional loss [Lee, 2013], we employ a memory module to compute a memory loss based on the cumulative class-level feature representation and model inference uncertainty aggregated all through the preceding training iterations. **(II)** Low computational cost: Only one network forward propagation is required to derive the additional loss term for the network by utilising a memory mechanism, as

opposed to more than one forward propagations required by other models [Rasmus et al., 2015; Miyato et al., 2016]. **(III)** Low consumption of memory footprint: Instead of storing all the predictions of all training samples in a large mapped file [Laine and Aila, 2017], our online updated memory module consumes very limited memory footprint, therefore potentially more scalable to training data of larger scale.

## 3.4 Experiments

We validate the effectiveness of MA-DNN on three widely adopted image classification benchmark datasets, with comparison to other state-of-the-art methods in Section 3.4.2 and ablation studies in Section 3.4.2.

### 3.4.1 Evaluation on Semi-Supervised Classification Benchmarks

**Datasets.** To evaluate our proposed MA-DNN, we select three widely adopted image classification benchmark datasets as detailed in the following.

**(1) SVHN** [Netzer et al., 2011]: A Street View House Numbers dataset including 10 classes (0∼9) of coloured digit images from Google Street View. The classification task is to recognise the central digit of each image. We use the format-2 version that provides cropped images sized at 32×32, and the standard 73,257/26,032 training/test data split.

**(2) CIFAR10** [Krizhevsky and Hinton, 2009]: A natural images dataset containing 50,000/10,000 training/test image samples from 10 object classes. Each class has 6,000 images with size 32×32.

**(3) CIFAR100** [Krizhevsky and Hinton, 2009]: A dataset (with same image size as CIFAR10) containing 50,000/10,000 training/test images from 100 more fine-grained classes with subtle inter-class visual discrepancy.

**Experimental Protocol.** Following the standard semi-supervised classification protocol [Kingma et al., 2014; Rasmus et al., 2015; Springenberg, 2016; Miyato et al., 2016], we randomly divide the training data into a small labelled set and a large unlabelled set. The number of labelled training images is 1,000/4,000/10,000 on SVHN/CIFAR10/CIFAR100 respectively, with the remaining 72,257/46,000/40,000 images as unlabelled training data. We adopt the common classification *error rate* as model performance measure, and report the average error rate over 10 random data splits.

**Implementation Details.** We adopt the same 10-layers CNN architecture as [Laine and Aila,

| Methods | SVHN | CIFAR10 | CIFAR100 |
|---|---|---|---|
| DGM* [Kingma et al., 2014] | 36.02 ± 0.10 | – | – |
| Γ-model [Rasmus et al., 2015] | – | 20.40 ± 0.47 | – |
| CatGAN* [Springenberg, 2016] | – | 19.58 ± 0.58 | – |
| VAT [Miyato et al., 2016] | 24.63 | – | – |
| ADGM* [Maaløe et al., 2016] | 22.86 | – | – |
| SDGM* [Maaløe et al., 2016] | 16.61 ± 0.24 | – | – |
| ImpGAN* [Salimans et al., 2016] | 8.11 ± 1.3 | 18.63 ± 2.32 | – |
| ALI* [Dumoulin et al., 2017] | 7.42 ± 0.65 | 17.99 ±1.62 | – |
| Π-model [Laine and Aila, 2017] | 4.82 ± 0.17 | 12.36 ± 0.31 | 39.19 ± 0.36 |
| Temporal Ensembling [Laine and Aila, 2017] | 4.42 ± 0.16 | 12.16 ± 0.24 | 37.34 ± 0.44 |
| Mean Teacher [Tarvainen and Valpola, 2017] | **3.95 ± 0.19** | 12.31 ± 0.28 | – |
| **MA-DNN (Ours)** | 4.21 ± 0.12 | **11.91 ± 0.22** | **34.51 ± 0.61** |

Table 3.1: Evaluation on semi-supervised image classification benchmarks in comparison to state-of-the-art methods. **Metric**: Error rate (%) ± standard deviation, **lower is better**. "–" indicates no reported result. "*" indicates generative models.

2017] (Table 3.2).

**Training details.** The convolutional neural network used for experiments on SVHN, CIFAR10 and CIFAR100 is detailed in Table 3.2. To train the network, we adopt the stochastic gradient descent algorithm with Nesterov momentum and set the batch size to 100, the weight decay to 0.0004. To compute the cross-entropy loss, we set a small label-smoothing factor of 0.001 in each class, such that the network prediction can better encode and reflect the model inference uncertainty [Pereyra et al., 2017]. The ratio of labelled/unlabelled samples is set to 50%/50% without tuning, which allows the memory module to receive a sufficient amount of labelled samples per batch for more reliably capturing the evolving feature space and the up-to-date model inference uncertainty. We set $\lambda$ in Eq. 6.7 to 1 to ensure equivalent importance of the supervised loss and the memory loss.

**Learning rate schedule.** On CIFAR10 and CIFAR100, the network is trained for 500 epochs with an initial learning rate of 0.1 and decayed linearly to 0 for the last 250 epochs. On SVHN, the network is trained for 200 epochs with an initial learning rate of 0.02 and decayed to 0 for the last 200 epochs.

| Layer | Descriptions |
|---|---|
| Input | $32 \times 32$ RGB Image |
| Conv1∼3 | $3 \times 3$ conv, 128 filters, LReLU (alpha=0.1), pad='same' |
| | $3 \times 3$ conv, 128 filters, LReLU (alpha=0.1), pad='same' |
| | $3 \times 3$ conv, 128 filters, LReLU (alpha=0.1), pad='same' |
| Pool1 | $2 \times 2$ max-pooling (dropout p=0.5), pad='same' |
| Conv4∼6 | $3 \times 3$ conv, 256 filters, LReLU (alpha=0.1), pad='same' |
| | $3 \times 3$ conv, 256 filters, LReLU (alpha=0.1), pad='same' |
| | $3 \times 3$ conv, 256 filters, LReLU (alpha=0.1), pad='same' |
| Pool2 | $2 \times 2$ max-pooling (dropout p=0.5), pad='same' |
| Conv7∼9 | $3 \times 3$ conv, 512 filters, LReLU (alpha=0.1), pad='valid' |
| | $1 \times 1$ conv, 256 filters, LReLU (alpha=0.1), pad='same' |
| | $1 \times 1$ conv, 128 filters, LReLU (alpha=0.1), pad='same' |
| Pool3 | average-pooling ($6 \times 6 \rightarrow 1$) |
| FC & Output | $128 \rightarrow$ number of classes ($K$) $\rightarrow$ $K$-way softmax |

Table 3.2: Network architecture. LReLU: LeakyReLU.

**Data augmentation.** We simply apply translation and colour jittering during training without any preprocessing prior to training. Additionally, horizontal flipping is applied on CIFAR10/100; whilst slight rotation is applied on SVHN.

**Comparison with state-of-the-art.** In Table 3.1, we compare our model to 11 state-of-the-art competitive methods with their reported results on SVHN, CIFAR10 and CIFAR100. Among all these methods, Mean Teacher is the only one that slightly outperforms our MA-DNN on the digit classification task. On the natural image classification tasks, our MA-DNN surpasses the best alternative (Temporal Ensembling) with a margin of 0.25%(12.16-11.91) and 2.83%(37.34-34.51) on CIFAR10 and CIFAR100 respectively. This indicates the potential performance superiority of the proposed MA-DNN in semi-supervised deep learning among various competitive semi-supervised learning algorithms, i.e. either performing on par with or outperforming the state of the art. Additionally, it can also be observed that MA-DNN outperforms more significantly on the more challenging dataset CIFAR100 with more fine-grained semantic structures among more classes. This suggests that the memory loss derived from the memory of model learning can en-

| Methods | SVHN | CIFAR10 | CIFAR100 |
|---|---|---|---|
| **Full** (ME+MND) | **4.21 ± 0.12** | **11.91 ± 0.22** | **34.51 ± 0.61** |
| **W/O** ME | 4.59 ± 0.11 | 12.63 ± 0.26 | 39.93 ± 0.34 |
| **W/O** MND | 6.75 ± 0.40 | 17.41 ± 0.15 | 41.90 ± 0.39 |

Table 3.3: Evaluation on the effect of individual memory loss terms. **Metric**: Error rate (%) ± standard deviation, **lower is better**. ME: Model Entropy; MND: Memory-Network Divergence.

hance more fine-grained class discrimination and separation to facilitate better semi-supervised learning. Therefore, MA-DNN is potentially more scalable than the other competitors on the image classification tasks that involve a larger number of classes.

**Computational Costs.** The per-batch distance computation complexity induced by memory assimilation and memory update is $\mathcal{O}(N_u K)$ and $\mathcal{O}(N_l)$ respectively, where $K$ is the number of memory slots, $N_l$ , $N_u$ are the numbers of labelled and unlabelled samples in each mini-batch. For computational efficiency, all the memory operations are implemented as simple matrix manipulation on GPU with single floating point precision. Overall, MA-DNN is computationally efficient in a number of ways: (i) Only one network forward propagation is required to compute the additional supervision signal, as opposed to more than one forward propagations required by Γ-model, VAT, Π-model and Mean-Teacher. (ii) The consumption of memory footprint is limited. The memory size of the memory module in MA-DNN is only proportional to the number of classes; while Temporal Ensembling requires to store the predictions of all samples in a large mapped file with a memory size proportional to the number of training samples. (iii) Unlike generative models including DGM, CatGAN, ADGM, SDGM, ImpGAN, and ALI, our MA-DNN does not need to generate additional synthetic images during training, therefore resulting in more efficient model training.

### 3.4.2 Ablation Studies and Further Analysis

**Effect of the Memory Loss.** We evaluate the individual contribution of two loss terms in the memory loss formulation (Eq. (3.7)): (1) the Model Entropy (ME) (Eq. (3.8)), and (2) the Memory-Network Divergence (MND) (Eq. (3.9)). We measure the impact of each loss term by the *performance drop* when removing it from the memory loss formulation.

Table 3.3 shows the evaluation results in comparison to the full memory loss formulation. We have the following observations: **(i)** Both loss terms bring positive effects to boost the model

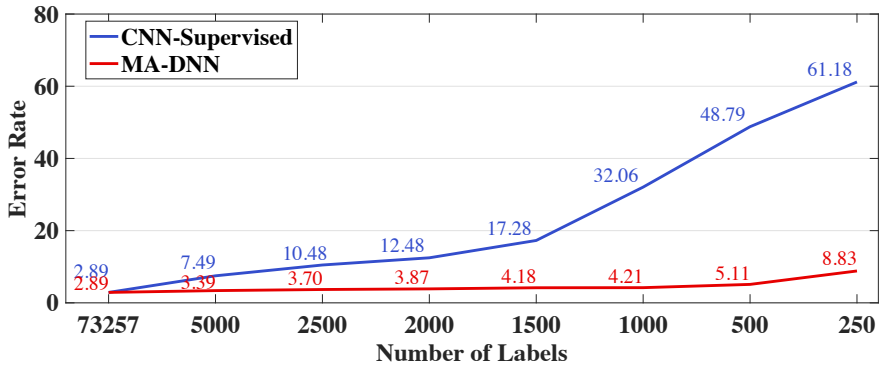



Figure 3.3: Evaluation on the robustness of the MA-DNN on varying number of labelled training samples. **Metric**: Error rate, **lower is better**.

performance. The classification error rates increase when either of the two loss terms is eliminated. **(ii)** The MND term effectively enhance the model performance. Eliminating the MND term causes performance drop of 2.54%(6.75-4.21), 5.50%(17.41-11.91), 7.39%(41.90-34.51) on SVHN, CIFAR10, and CIFAR100 respectively. This indicates the effectiveness of encouraging the network predictions to be consistent with reliable memory predictions derived from the memory of model learning. **(iii)** The ME term is also effective. Eliminating the ME term causes performance drop of 0.38%(4.59-4.21), 0.72 %(12.63-11.91), 5.42%(39.93-34.51) on SVHN, CIFAR10, and CIFAR100 respectively. This suggests the benefit of penalising class distribution overlap and enhancing class separation, especially when the amount of classes increase – more classes are harder to be separated. Overall, the evaluation in Table 3.3 demonstrates the complementary joint benefits of the two loss terms to improve the model performance in semi-supervised deep learning.

**Labelled Training Sample Size.** We evaluate the robustness of MA-DNN over varying numbers of labelled training samples. We conduct this evaluation on SVHN by varying the number of labelled samples from 73,257 (all training samples are labelled) to 250. As comparison, we adopt the supervised counterpart *CNN-Supervised* trained only using the same labelled data without the memory module. Figure 3.3 shows that as the size of labelled data decreases, the model performance of CNN-Supervised drops from 61.18% (given 73,257 labelled samples) to 2.89% (given 250 labelled samples), with a total performance drop of 58.29% in error rate. In contrast, the performance of MA-DNN degrades only by 5.94%(8.83-2.89). This indicates the proposed MA-DNN can effectively leverage additional unlabelled data to boost the model performance when both small-sized labelled and large-sized unlabelled training data are provided.

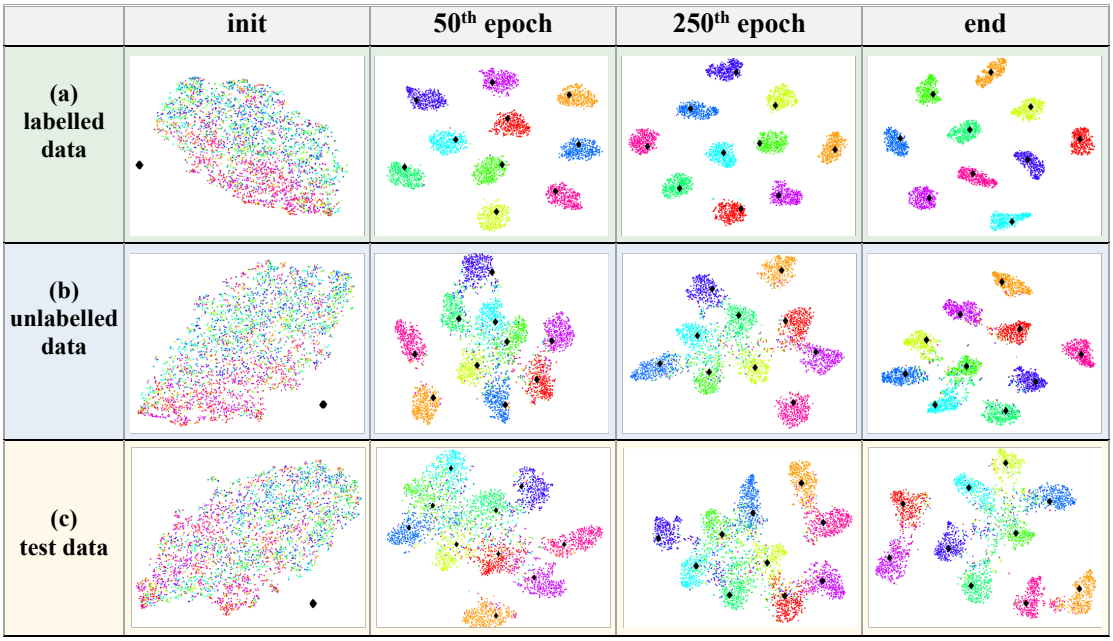



Figure 3.4: Visualisation on the evolution of key embeddings (denoted as the *black dots*) and the multi-class data distribution (denoted as dots in colours) of **(a)** labelled data, **(b)** unlabelled data, **(c)** test data from CIFAR10 in the feature space during training. Data projection in 2-D space is attained by tSNE [Maaten and Hinton, 2008] based on the feature representation extracted on the *same* sets of data using the CNN at different training stages.

**Evolution of the Memory Module.** As aforementioned, the two types of class-level memorisable information recorded in the memory module is **(1)** the class-level feature representation (key embeddings), and **(2)** the model inference uncertainty (value embeddings). To understand how the memory module is updated during training, we visualise the evolution of the key embeddings and value embeddings in Figure 3.4, 3.5 and qualitatively analyse their effects as below.

**Effect of the Key Embeddings.** As Figure 3.4 shows, the key embeddings (denoted as the *black dots*) are essentially updated as the cluster centroids to capture the global manifold structure in the feature space. In particular, we have the following observations: **(i)** Figure 3.4(a) shows that although the key embeddings are initialised as **0** without imposing prior knowledge, they are consistently updated to capture the underlying global manifold structure of the labelled data after training for certain period (50 epochs). **(ii)** Figure 3.4(b) shows that although there is severe class distribution overlap in the feature space initially, the class distribution overlap of unlabelled data tends to be gradually mitigated as the model is trained. **(iii)** Figure 3.4(c) shows that the key embeddings also roughly capture the global manifold structure of the unseen test data, even though the network is not optimised to fit towards the test data distribution. Overall, these observations are in line with our motivation of recording the accumulatively updated cluster centroids

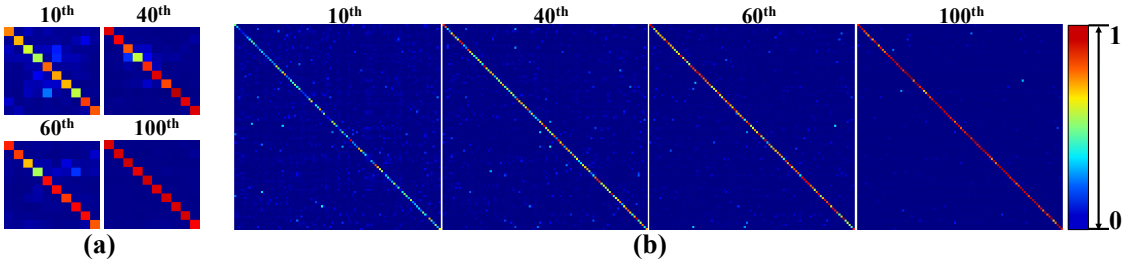



Figure 3.5: Visualisation on the evolution of value embeddings on **(a)** CIFAR10 and **(b)** CIFAR100. In each block, each row corresponds to a per-class value embedding, i.e. a multi-class probabilistic prediction that encodes the class-level network inference uncertainty at different epochs during training.

as the key embeddings for deriving the probabilistic assignments on unlabelled samples based on the *cluster assumption*. Moreover, the evolution of unlabelled data distribution also qualitatively suggests that our memory loss serves to penalise the class distribution overlap and render the class decision boundaries to lie in the low density region. Note that the 2D visualisation of high-dimensional data may not perfectly reflect how classes are separated.

**Effect of the Value Embeddings.** As Figure 3.5 shows, the value embeddings essentially record the model inference uncertainty at the class level. At the initial training stages, the value embeddings reflect much higher inference uncertainty (multimodal distribution with higher entropy), but progressively reflect much lower inference uncertainty (peaked distribution with lower entropy) as the model is trained. In fact, when removing the value embeddings, the probabilistic assignments on unlabelled samples can become particularly unreliable at the earlier training stages, which even leads to performance drops of 0.69/1.94/2.78% on SVHN/CIFAR10/100. Hence, the value embeddings can serve to reflect the class separation in the label space, and be utilised to smooth the probabilistic assignments with uncertainty for deriving more reliable memory predictions.

**Evolution of Memory Predictions.** We visualise the evolution of memory predictions on the unlabelled samples from CIFAR10 at different training stages in Figure 3.6. It can be observed that the memory predictions are progressively improving from more uncertain (ambiguous) to more confident on the unlabelled training samples. This not only demonstrates the good convergence property of the MA-DNN, but also indicates how the memory loss takes effect in model learning – (1) penalising class distribution overlap when given uncertain memory predictions at the earlier training stages while (2) encouraging the network predictions to be consistent with confident memory predictions, such that the unlabelled data is fitted optimally towards the underlying

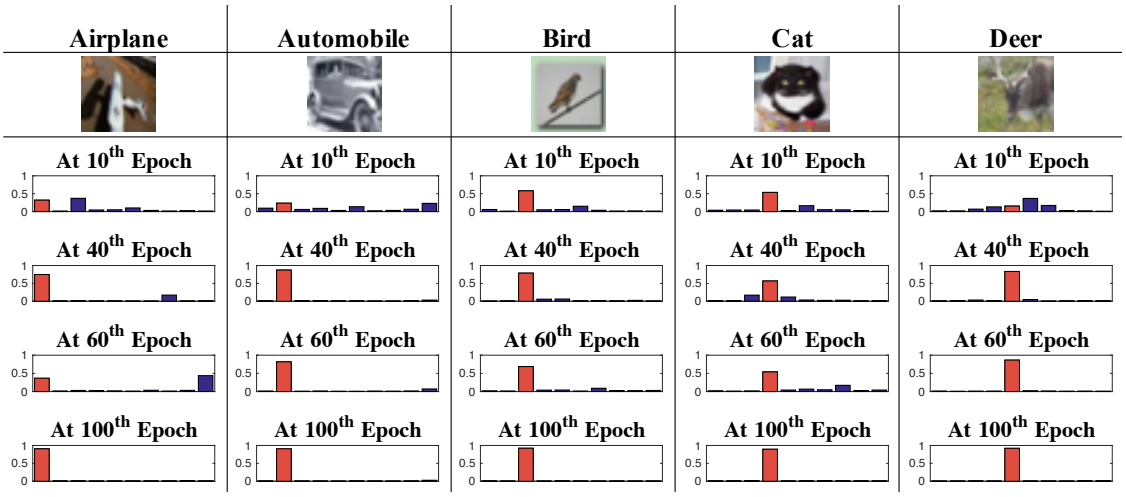



Figure 3.6: Evolution of memory predictions of randomly selected *unlabelled* samples from CIFAR10. The *Red* bar corresponds to the *missing* ground-truth class.

manifold structure.

## 3.5 Summary

In this work, we presented a novel Memory-Assisted Deep Neural Network (MA-DNN) to enable semi-supervised deep learning on sparsely labelled and abundant unlabelled training data. The MA-DNN is established on the idea of exploiting the memory of model learning to more reliably and effectively learn from the unlabelled training data. In particular, we formulated a novel assimilation-accommodation interaction between the network and an external memory module capable of facilitating more effective semi-supervised deep learning by imposing a memory loss derived from the incrementally updated memory module. Extensive comparative evaluations on three semi-supervised image classification benchmark datasets validate the advantages of the proposed MA-DNN over a wide range of state-of-the-art methods. We also provided detailed ablation studies and further analysis to give insights on the model design and performance gains.

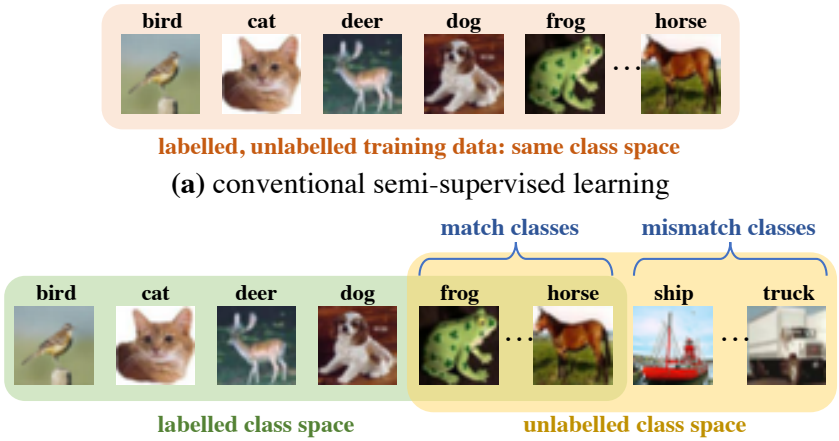
# Chapter 4

# Open-Set Semi-Supervised Learning by Uncertainty-Aware Self-Distillation

---

Recent developments [Kingma et al., 2014; Rasmus et al., 2015; Miyato et al., 2016; Sajjadi et al., 2016; Laine and Aila, 2017; Tarvainen and Valpola, 2017; Athiwaratkun et al., 2019; Berthelot et al., 2019] have pushed the limit of SSL drastically, leading to increasingly generalisable DNNs. With a substantial fraction of labels discarded, recent advanced methods [Laine and Aila, 2017; Tarvainen and Valpola, 2017; Athiwaratkun et al., 2019] can even approach the performance of fully supervised learning. Most of the prior works in semi-supervised learning [Sajjadi et al., 2016; Laine and Aila, 2017; Tarvainen and Valpola, 2017; Miyato et al., 2018; Chen et al., 2018b; Athiwaratkun et al., 2019; Wang et al., 2019; Berthelot et al., 2019], however, focus on a *close-set* semi-supervised learning setting, where the unlabelled data samples are assumed to lie in the same label space as the labelled data. The best results on close-set semi-supervised image classification benchmarks are mostly achieved by consistency regularisation, which generally enforces the distributional smoothness under randomisation in input images or model weights. For instance, Temporal Ensembling [Laine and Aila, 2017], Mean Teacher [Tarvainen and Valpola, 2017] are two representative techniques that keep an *exponential moving average* (EMA) in output or weight space to derive an unsupervised consistency cost on unlabelled data.

Built upon a *de facto* artificial assumption that labelled and unlabelled training data are drawn from an identical class space (i.e. every unlabelled sample must belong to one of the known classes), existing SSL methods are not practically deployable and scalable in practice. In fact,




**(a)** conventional semi-supervised learning

**(b)** realistic semi-supervised learning

Figure 4.1: **(a)** In conventional semi-supervised learning, both labelled and unlabelled training data come from an identical class distribution. **(b)** In real-world scenario, however, class distribution mismatch often exits between the labelled and unlabelled data.

as illustrated by [Oliver et al., 2018], the model performance of existing representative semi-supervised learning techniques would degrade significantly when tested in an *open-set* semi-supervised learning setting, where there exists class distribution mismatch between the labelled and unlabelled sets. This is because, unlabelled data are unlikely to be manually purified beforehand in many real-world applications for meeting the *de facto* assumption. More probably, unlabelled data are sampled from a class distribution with an *unknown* class mismatch rate against the labelled class distribution. Lacking algorithmic consideration to deal with the class distribution mismatch between labelled and unlabelled data, state-of-the-art SSL algorithms generally suffer severe performance degradation when deployed to such realistic settings.

In this work, we investigate the more realistic and under-studied semi-supervised learning scenario with *class distribution mismatch* between limited labelled and abundant unlabelled data sets. In particular, unlike the conventional SSL setting, we consider the unlabelled data is drawn from a mixture of known and unknown classes (Figure 4.1). This new problem poses a unique research question mostly ignored in existing SSL literature: *How can we maximise the value of any relevant unlabelled data, given no prior knowledge about whether an unlabelled sample belongs to a known class?*

Compared to conventional SSL, the challenge of realistic SSL is partly due to lacking the separation of known and unknown classes on the unlabelled training data. Together with the notorious ***overconfidence*** issue of deep neural networks (DNNs) [Nguyen et al., 2015], it is not surprised that contemporary SSL methods can easily produce corrupted, *overconfident* un-

supervised learning signals that incur catastrophic error propagation. For instance, in *entropy minimisation* for SSL [Grandvalet and Bengio, 2005; Lee, 2013], model predictions are blindly enforced to be "*confident*" (i.e. low-entropy) on unlabelled samples, despite that these samples may be unrelated to the target learning task at hand. In *consistency regularisation* [Tarvainen and Valpola, 2017], the inherent *overconfident* tendency in DNNs can also reinforce the wrong class assignments of those irrelevant unlabelled samples to the known classes. Therefore, to exploit unconstrained unlabelled data effectively, we address this realistic SSL problem based on two essential algorithmic considerations: (1) self-discover and discard irrelevant unlabelled data on-the-fly; and (2) formulate reliable learning signals that avoid overconfident class assignments.

Specifically, we formulate a generic and novel SSL deep learning algorithm, named **Uncertainty-Aware Self-Distillation** (UASD), which addresses the aforementioned challenge in a systematic end-to-end formulation. Our model formulation UASD is partially inspired by recent works in **out-of-distribution (OOD) detection** – a task of detecting OOD samples. An intuitive approach in OOD is to identify OOD samples based on confidence scores estimated as the maximum softmax probabilities [Hendrycks and Gimpel, 2017]. However, softmax-based confidence estimate by a *single* DNN can be problematic, as DNNs generally suffer from *overconfidence* [Nguyen et al., 2015], e.g. Feeding random noise to a DNN can give rise to a maximal probability score over 99.6%. To address this, one line of research focuses on *confidence calibration* [Liang et al., 2018; Lee et al., 2018b; DeVries and Taylor, 2018; Hendrycks et al., 2019] to form *softer* predictive distributions that encompass uncertainty. Rooted in similar spirit as these prior works, we consider *soft targets* can serve as an indicator for OOD detection. In particular, we introduce a novel simple OOD filter to automatically discard OOD samples on-the-fly by comparing the confidence scores of training samples to a referential confidence score derived from a validation set. Compared to existing approaches in OOD detection, our approach does not require heavy computation cost to train an OOD filter, nor the need of auxiliary OOD training samples.

Besides discarding OOD samples on-the-fly, UASD is also specialised in forming the *soft targets* that could further serve as *regularisers* to empower more robust semi-supervised learning under class distribution mismatch. Critically, UASD prevents the tendency of *overconfidence* in DNN, a fundamental limitation that existing SSL methods commonly suffer – consequently causing their error propagation and catastrophic degradation in the more realistic SSL setting. This is achieved by formulating a sequence of ensemble models aggregated accumulatively on-

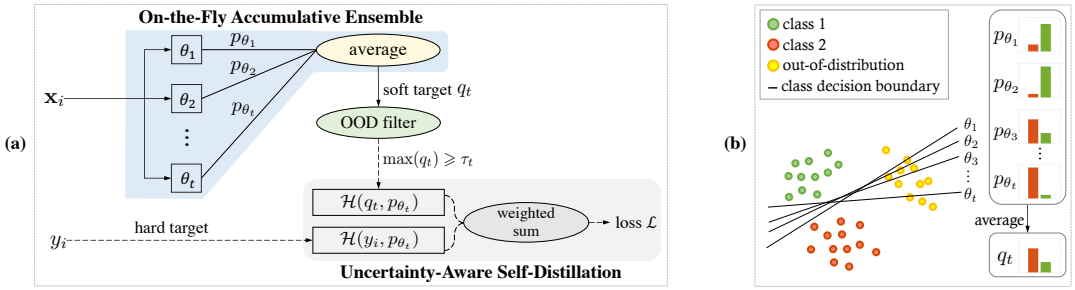



Figure 4.2: **(a) Approach overview**: The predictions from historic stochastic passes on each sample are accumulatively *averaged* to derive a smooth predictive distribution $q_t$. For robust SSL, $q_t$ is used for unlabelled training data filtering and uncertainty-aware self-distillation. **(b) Schematic illustration**: If a sample is not consistently assigned to a class by an ensemble of classifiers, it is likely to lie out-of-distribution.

the-fly for joint *Self-Distillation* and *OOD filtering*. Unlike existing SSL methods that derived their overconfident learning signals based on the *de facto* assumption, our formulation is aware of the uncertainty of whether an unlabelled sample likely lies in- or out-of-distribution, and selectively learns from the unconstrained unlabelled data.

In summary, our **contribution** is three-fold:

- We study semi-supervised learning under class distribution mismatch – a realistic SSL scenario largely ignored in existing SSL literature. To our knowledge, this work is the first attempt to systematically address this new problem.
- We formulate a novel algorithm, *Uncertainty-Aware Self-Distillation* (UASD), for solving the unique SSL challenge involved in class distribution mismatch. UASD overcomes the *overconfident* issue of DNNs and enables robust SSL under class distribution mismatch.
- We provide extensive benchmarking results in this realistic SSL scenario, including our proposed UASD and *six* representative state-of-the-art SSL methods on *three* image classification datasets: CIFAR10, CIFAR100 and TinyImageNet. Remarkably, UASD outperforms all the strong competitors often by large margins, and demonstrates great potential to exploit the unconstrained unlabelled data.

## 4.1 Uncertainty-Aware Self-Distillation

We consider a realistic semi-supervised learning (SSL) scenario with class distribution mismatch, where we have access to a limited amount of labelled samples $\mathcal{D}_l = \{\mathbf{x}_{i,l}, y_i\}_{i=1}^{N_l}$, and abundant unlabelled samples $\mathcal{D}_u = \{\mathbf{x}_{i,u}\}_{i=1}^{N_u}$. Each labelled sample $\mathbf{x}_{i,l}$ belongs to one of $K$

known classes $\mathcal{Y} = \{y_k\}_{k=1}^{K}$, while any unlabelled sample $\mathbf{x}_{i,u}$ is **not** guaranteed to be one of these $K$ known classes. The class distribution mismatch proportion between $\mathcal{D}_l$ and $\mathcal{D}_u$ is also **unknown**. Our ultimate goal is to exploit useful unlabelled data to boost the learning task at hand. Compared to conventional SSL [Chapelle et al., 2009], this scenario raises a unique challenge on how to mitigate the risk of error propagation mostly incurred by overconfidently assigning out-of-distribution samples to the known classes.

### 4.1.1 Approach Formulation

To perform SSL under class distribution mismatch, we need to achieve two goals concurrently: (1) minimise the negative influence incurred by irrelevant unlabelled data; and (2) maximise the exploitation of relevant unlabelled data to improve the target learning task. To this end, we propose **Uncertainty-Aware Self-Distillation** (UASDx), a unified SSL algorithmic framework (as shown in Figure 4.2) that jointly perceives data ambiguity with predictive uncertainty, and produces soft targets as effective regularisers to selectively learn from unlabelled data with mismatched classes.

### 4.1.2 On-the-Fly Accumulative Ensemble

To formulate UASD, we exploit the generic model ensemble principle [Schapire, 1990; Breiman, 2001], with an aim to sufficiently reduce model misspecification and yield soft targets as regularisers. The rationale is that a committee of models can cover different regions of the version space [Mitchell, 1982], as different models tend to make predictions and mistakes differently. Thus, by aggregating predictions from multiple models, we can not only derive smoother predictive distributions (a.k.a. *soft targets*) that encompass predictive uncertainty, but also produce a stronger model that offers richer knowledge beyond the available class label information. However, training a cumbersome ensemble is computationally expensive. To address this issue, we construct the ensemble model on-the-fly by an accumulation strategy.

Specifically, we exploit a sequence of ensemble models that accumulatively grows the ensemble size on-the-fly. Formally, at the $t$-th epoch we build an ensemble by aggregating all the historic networks $\{\theta_j\}_{j=0}^{t}$. Given a sample $\mathbf{x}_i$, we derive its ensemble prediction $q_t(y|\mathbf{x}_i)$ by averaging over all the preceding network predictions:

$$q_t(y|\mathbf{x}_i) = \frac{1}{t}\sum_{j=0}^{t-1} p(y|\mathbf{x}_i;\theta_j) \tag{4.1}$$

where $p(y|\mathbf{x}_i;\theta_j)$ denotes the network prediction at the $j$-th epoch. It is worth noting that, the stochasticity induced by various data augmentation, batch norm, and network weights in different stochastic passes enables building an ensemble out of an increasing amount of models with diverse decision boundaries. Joining with the training process, we can easily scale up the ensemble size by modulating the network aggregating frequency. Crucially, averaging all historic predictions helps to reduce the bias by cancelling out mistaken and overconfident class assignments made by individual networks. This effectively produces smoother ensemble predictive distributions – a type of *soft targets* that naturally encompass both the predictive uncertainty [Lakshminarayanan et al., 2017] and regularities [Hinton et al., 2015] discovered by a very large ensemble of models.

### 4.1.3 Unlabelled Training Data Filtering

For robust SSL under class distribution mismatch, we leverage the soft targets $q_t$ derived in Eq. (4.1) as an indicator to discard the potentially irrelevant unlabelled training data. Since $q_t$ reflects the agreement among the historic networks in a frequentist perspective [Dawid, 1982], its maximal class probability indicates the best consensus a sample is assigned to a specific class. Accordingly, we define the predictive confidence score on each sample as:

$$c_t(\mathbf{x}_i) = \max(q_t(y|\mathbf{x}_i)) \tag{4.2}$$

where a lower confidence score $c_t(\mathbf{x}_i)$ reflects higher *predictive uncertainty*, indicating the sample is *likely* to lie out-of-distribution (OOD) and uncorrelated to the core learning task at hand. To minimise the harmful effect incurred by irrelevant unlabelled samples, we define an OOD filter to discard the samples with low confidence scores:

$$f(\mathbf{x}_i;\tau_t) = \begin{cases} 1, & \text{if } c_t(\mathbf{x}_i) \geqslant \tau_t, \text{ selected} \\ 0, & \text{if } c_t(\mathbf{x}_i) < \tau_t, \text{ rejected} \end{cases} \tag{4.3}$$

where $f(\mathbf{x}_i;\tau_t)$ specifies a batch-wise binary sample filtering criterion to select samples for model learning based upon a confidence threshold $\tau_t$. The threshold $\tau_t$ is often heuristically set, which however, is unsuitable in our context, as it depends heavily on the in-training model with high dynamics. Thus, we dynamically estimate $\tau_t$ in a data-driven manner by using the validation set (10% of training data) of known classes as reference. Formally, we compute $\tau_t$ as the *average* confidence score on the in-distribution validation samples, and refresh $\tau_t$ iteratively per epoch

during the course of training. It is worth noting that while our aim is to discard out-of-distribution samples, certain amount of in-distribution samples may also be discarded. This suggests a *trade-off* between maximising the usage of unlabelled data and minimising the risk of error propagation induced by out-of-distribution samples. Since our confidence scores have the less overconfident tendency owing to ensembling multiple network predictions for each sample, the estimated confidence scores can be better utilised to delimit between the in- and out-of-distribution samples, thus enabling a better trade-off to incorporate more useful unlabelled data for model training.

## 4.2 Model Optimisation

To enable semi-supervised learning, we employ the soft target $q_t$ to derive a self-supervision signal, which is imposed as a regulariser to learn additionally from the relevant unlabelled data, i.e. samples likely to be correlated to the learning task. Specifically, we capitalise the rich information encoded in soft targets for model learning, including (1) the regularities among known classes, and (2) the predictive uncertainty. All such information are discovered by on-the-fly accumulative ensembling *without* the need of class labelling. Therefore, the soft targets naturally serve to propagate soft class assignments on the unlabelled data in an *unsupervised* manner.

### 4.2.1 Semi-Supervised Learning Objective

Formally, motivated by the generic distillation principle [Hinton et al., 2015], we consider the soft targets as a kind of teaching signal and formulate the final SSL objective with uncertainty-aware self-distillation as:

$$\mathcal{L} = \mathcal{H}(y_{\text{true}}, p_\theta) + w(t) f(\cdot;\tau_t) \cdot \mathcal{H}(q_t, p_\theta) \tag{4.4}$$

where the first term refers to the standard *supervised* cross-entropy loss, computed between the network prediction $p_\theta$ and the ground-truth labels $y_{\text{true}}$. The second term is the *unsupervised* uncertainty-aware self-distillation loss, computed as the cross-entropy between $p_\theta$ and the soft targets $q_t$. The OOD filter $f(\cdot;\tau_t)$ (Eq (4.3)) is aware of uncertainty and used to discard the potentially irrelevant samples of low confidence scores. In the beginning of training, the soft targets may not be sufficiently informative due to lacking diversity in ensembling and reliability in predictions. Thus, for robust model optimisation, we utilise a ramp-up weighting function $w(t)$ to gradually increase the importance of the self-distillation loss.

### 4.2.2 Model Training

An algorithmic overview is summarised in Algorithm 2.

---

**Algorithm 2** Uncertainty-Aware Self-Distillation (UASD)

---

**Require:** Labelled data $\mathcal{D}_l = \{\mathbf{x}_{i,l}, y_i\}_{i=1}^{N_l}$. Unlabelled data $\mathcal{D}_u = \{\mathbf{x}_{i,u}\}_{i=1}^{N_u}$.

**Require:** Trainable neural network $\theta$. Ramp-up weighting function $w(t)$.

**for** $t = 1$ **to** *max_epoch* **do**

Refresh confidence threshold $\tau_t$ *per* epoch.

**for** $k = 1$ **to** *max_iter_per_epoch* **do**

Forward propagation to accumulate network prediction $q_t(y|\mathbf{x}_i)$ (Eq (4.1)) for every in-batch sample.

Apply OOD filtering (Eq (4.2), (4.3)).

Update network parameters $\theta$ with loss function Eq (4.4).

**end for**

**end for**

---

## 4.3 Discussion

We consider the failure of existing methods in addressing *open-set* semi-supervised learning is caused by their general tendency of producing *overconfident* class assignments on unlabelled data, regardless of the underlying class distribution. In fact, they are likely to spread the wrong class labels to unlabelled samples lying out-of-distribution. To resolve this issue, we propose to accumulatively aggregate the predictions from a growing amount of networks by *equal averaging*, thus leading to much *softer* class assignments for more reliable SSL under class distribution mismatch. Overall, our approach has several unique merits to benefit SSL under class distribution mismatch: **(I)** Instead of computing the ensemble predictions based on *exponential moving average* in the *logit space* as Temporal Ensembling [Laine and Aila, 2017], our ensemble predictions are acquired by accumulative ensembling, which keeps an *equal average* in the *prediction space* to derive *less overconfident* and softer class assignments. **(II)** Rather than filtering the unlabelled data using an OOD detector pre-trained on OOD samples, we leverage the *soft targets* as an indicator to discard OOD samples on-the-fly, therefore eschewing the requirement of OOD training data and additional training cost for an OOD filter. **(III)** We integrate *Self-Distillation*

and *OOD filtering* in a unified end-to-end training framework, which allows the model to benefit learning from the unconstrained unlabelled data in a more reliable way.

Our proposed approach is also closely related to **knowledge distillation** [Hinton et al., 2015], where the goal is to transfer the knowledge from an ensemble of multiple DNNs into a single DNN. Typically, a *student* network is supervised by the soft targets generated by averaging the network outputs from a cumbersome whole ensemble of *teacher* networks. Compared to the ground truth one-hot labels, i.e. hard targets, soft targets are smoother predictive distributions with higher entropy that encode more information discovered by multiple models. Inspired by the recently proposed *online distillation* [Anil et al., 2018; Lan et al., 2018], our approach particularly aims to exploit the knowledge discovered *on-the-fly* by accumulating predictions of all historic stochastic forward passes. This yields *soft targets* that encode uncertainty for OOD data filtering; and more importantly, produce *less overconfident* and softer class assignments on unlabelled data samples to avoid catastrophic error propagation. Although *self-distillation* has been proposed previously by Zhang et al. [Zhang et al., 2019], our model formulation fundamentally differs from this prior work. In particular, we focus on deriving soft targets that encompass model inference uncertainty by aggregating predictions from *one* classifier, while the prior work trains *multiple* classifiers at different network layers to obtain the ensemble predictions more efficiently.

Our proposed ensemble strategy is also related to *Monte Carlo Dropout* [Gal and Ghahramani, 2016] (MC-Dropout), a Bayesian approximation of a well known probabilistic model: the Gaussian process. However, the ensemble predictions are derived very differently between our model and the MC-Dropout. In MC-Dropout, ensembled network predictions are derived by averaging predictions from the same networks under different dropout, which means the different networks share very similar model weights. In contract, in our ensemble strategy, networks from different training iterations are quite diverse, which yield ensembled predictions that are less prone to be overconfident.

## 4.4 Experiments

**Implementation details.** For a comprehensive and fair comparison, our experiments are built upon the open-source Tensorflow implementation by Oliver et al. [Oliver et al., 2018]. It uses the standard Wide ResNet [Zagoruyko and Komodakis, 2016], i.e. *WRN-28-2*, as the base network and Adam optimiser [Kingma and Ba, 2014] for training. We revise the default 10-dimensional

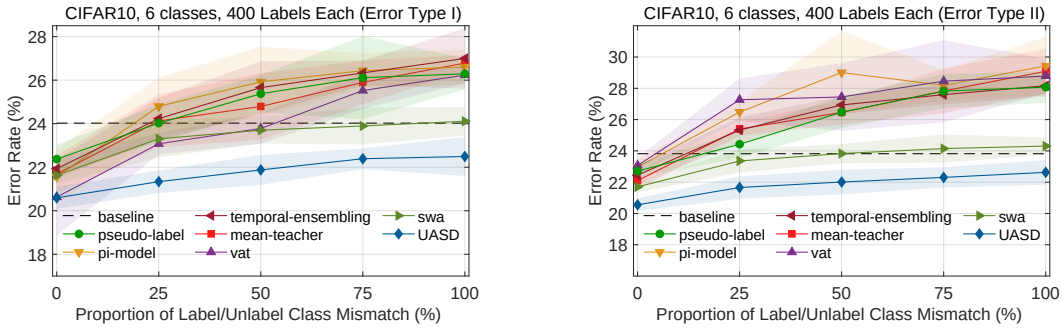



Figure 4.3: Experiment results of different SSL methods on CIFAR10 under varying class distribution mismatch proportion. **Left** (I): Test error rates are reported at the point of lowest validation error. **Right** (II): Test error rates are reported as the median of last 20 epochs. Shaded area indicates the standard deviation over five runs. Tabular results are provided in Table 4.5 and 4.7.

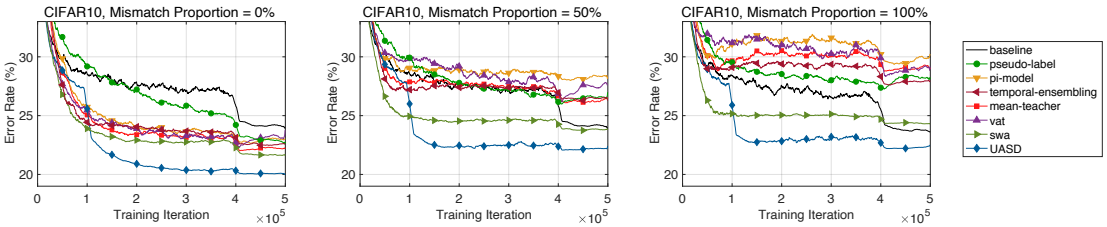



Figure 4.4: Smoothed learning curves averaged over five runs of different SSL methods on CIFAR10. **Left/Middle/Right** correspond to learning curves under class distribution mismatch proportion of 0/50/100%.

classification layer to $K$-dimension, where $K$ is the number of known classes in the labelled data. Unless stated otherwise, all hyper-parameters, the ramp-up function, and training procedures are the same as that of [Oliver et al., 2018]. For all comparisons, we report the supervised baseline results using only labelled data for training. More details of implementation are given in Table 4.2 and 4.3.

**Datasets.** We use three image classification benchmark datasets. **(1)** CIFAR10: A natural image dataset with 50,000/10,000 training/test samples from 10 object classes. **(2)** CIFAR100: A dataset of 100 fine-grained classes, with the same amount of training/test samples as CIFAR10. **(3)** TinyImageNet: A subset of ImageNet [Deng et al., 2009] with 200 classes, each of which has 500/50 training/validation images. For all datasets, we resize the images to 32×32.

**Compared methods.** We compare our method with six representative state-of-the-art SSL methods, including **(1)** pseudo-labels [Lee, 2013], **(2)** VAT [Miyato et al., 2016], **(3)** Π-model [Sajjadi et al., 2016; Laine and Aila, 2017], **(4)** Temporal Ensembling [Laine and Aila, 2017], **(5)** Mean-Teacher [Tarvainen and Valpola, 2017], and **(6)** SWA [Athiwaratkun et al., 2019]. All these methods introduce an additional unsupervised supervision signal, originally proposed and tested

under the conventional SSL setting *without* class distribution mismatch. To preserve the nature of these methods, we replicate them following the procedures as [Oliver et al., 2018] – all share the same network architecture, data augmentation, optimiser and training time. We also compare all the methods to a baseline that is trained with only labelled data. All the other methods are trained on the labelled data and unlabelled data with in- and out-of-distribution samples. In all experiments, we test each method for five runs with a same set of random seeds to choose the labelled samples. We report the averaged error rate with mean and standard deviation over five runs.

### 4.4.1 Evaluation on CIFAR10

**Evaluation protocol.** To simulate more realistic SSL with class distribution mismatch, we construct the unlabelled data with unknown classes not present in the labelled data. Following [Oliver et al., 2018], we perform experiments on CIFAR10 for a 6-class classification task, using *400 labels per class*. The labelled set contains 6 classes of animals: *bird, cat, deer, dog, frog, horse*; while the unlabelled data comes from 4 classes, with a varying class distribution mismatch proportion from 0% to 100%. For instance, for a mismatch proportion of 50%, the unlabelled data contains classes of *airplane, automobile, frog, horse*. The test errors are reported on the 6 known classes. More details about the evaluation protocol are given in Table 4.4.

**Evaluation results.** Figure 4.3 shows experiment results on CIFAR10, including six SSL methods and our UASD under varying class distribution mismatch proportion. The two diagrams (left and right) show the test error rates in two ways: (I) test error rate at the point of the lowest validation error; (II) median test error rate of last 20 epochs. It can be observed that when increasing the amount of unlabelled samples from unknown classes, the performance of most state-of-the-art SSL methods degrade drastically, except SWA [Athiwaratkun et al., 2019], a very recent SSL method that performs weight averaging during training.

Compared to SWA, UASD surprisingly improves the error rates and suffers much less degradation under high mismatch proportions – which indicates its capability to exploit unlabelled data in a more reliable way. Moreover, the error rates of UASD stay consistently in two ways of test error calculation, whilst other methods show more severe performance degradation when reporting the median error rate in the last 20 epochs. This means the other methods commonly suffer unstable degradation at the end of training, while UASD exhibits much more robust convergence.

| Method | CIFAR100 | TinyImageNet | CIFAR100 + TinyImageNet |
|---|---|---|---|
| baseline | $39.79 \pm 1.19$ | $61.64 \pm 0.59$ | $48.31 \pm 0.63$ |
| pseudo-label | $43.30 \pm 0.57$ | $62.41 \pm 0.57$ | $53.3 \pm 0.73$ |
| VAT | $43.78 \pm 1.15$ | $63.75 \pm 0.69$ | $50.55 \pm 0.55$ |
| Π-Model | $42.96 \pm 0.46$ | $61.79 \pm 0.67$ | $53.05 \pm 2.21$ |
| Temporal Ensembling | $41.27 \pm 0.76$ | $60.69 \pm 0.31$ | $47.88 \pm 0.64$ |
| Mean-Teacher | $40.98 \pm 0.98$ | $\mathbf{60.54 \pm 0.31}$ | $49.67 \pm 1.95$ |
| SWA | $\mathbf{37.66 \pm 0.48}$ | $\mathbf{57.97 \pm 0.42}$ | $\mathbf{44.61 \pm 0.52}$ |
| Ours | $\mathbf{35.93 \pm 0.60}$ | $\mathbf{57.15 \pm 0.76}$ | $\mathbf{42.83 \pm 0.25}$ |

Table 4.1: Results on CIFAR100 and TinyImageNet averaged over 5 runs. Results with reduction in error rate compared to baseline are highlighted in **bold**. Best results are highlighted in **red**.

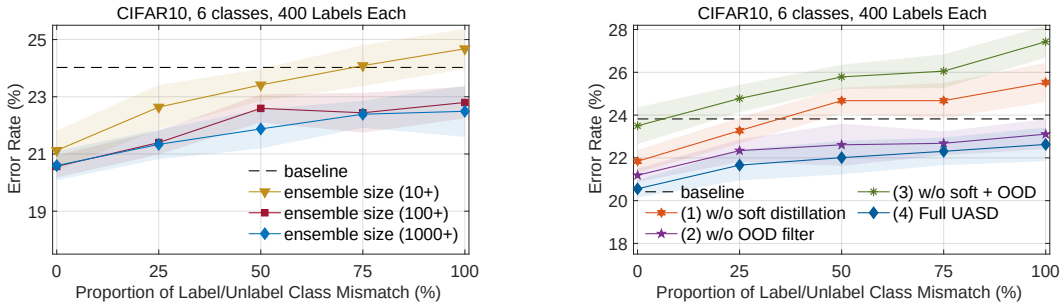



Figure 4.5: Ablative evaluation (test error rates on CIFAR10). **Left**: Ensemble size analysis. **Right**: Loss formulation analysis. Tabular results are provided in Table 4.9 and 4.10.

**Learning dynamics analysis.** To understand the learning dynamics, we visualise the learning curves in terms of test error rate during training in Figure 4.4. It is evident that UASD remains superior learning performance compared to other SSL methods under different class distribution mismatch proportions, demonstrating more stable convergence and more reduction in error rate. The relative benefits compared to other SSL methods are more significant under higher class mismatch proportions, e.g. 50%, 100%. This suggests that UASD does yield more reliable supervision signals to guarantee the effectiveness and robustness of SSL under class distribution mismatch.

### 4.4.2 Evaluation on CIFAR100 and TinyImageNet

**Evaluation protocols.** We conduct experiments on CIFAR100 and TinyImageNet to evaluate SSL under large class distribution mismatch in larger class space, including three settings as described next.

- On CIFAR100, we use the first half classes (1-50) as labelled classes, and the 25-75 classes as unlabelled classes, leading to a class distribution mismatch proportion of **50%** between labelled and unlabelled data.
- On TinyImageNet, we use the 1-100 classes as labelled classes and the 50-150 classes as unlabelled classes, which results in a mismatch proportion of **50%**.
- We further test in a **cross-dataset** scenario using 100 classes from CIFAR100 as the labelled classes, and 200 classes from TinyImageNet as the unlabelled classes, which gives a mismatch proportion of **86.5%**.

For all experiments in this section, we use *100 labels per class* and report the test error rate as the median of last 20 epochs to reflect the final convergence.

**Evaluation results.** Table 4.1 shows UASD remains remarkably better than other methods when learning under large class distribution mismatch in the finer-grained classification tasks on CIFAR100 and TinyImageNet. While most SSL methods suffer model degradation, UASD consistently outperforms all of them in all settings. It improves upon the supervised baseline with test error reduction of $3.86\%, 4.49\%, 5.48\%$. Crucially, it succeeds even when a large class distribution mismatch proportion (i.e. 86.5%) exists across two datasets (CIFAR100 + TinyImageNet). This shows the efficacy of UASD in exploiting unconstrained unlabelled data coming from *unknown but related classes*, or even *unseen distribution of another dataset*.

### 4.4.3 Ablative Analysis

To assess different aspects in our algorithmic formulation, we conduct ablative evaluation by changing one individual factor at a time whilst keeping others fixed.

**(I) Ensemble size.** As aforementioned, the ensemble size is accumulatively growing on-the-fly, which results in a very large ensemble in the end of training (e.g. 1000+ on CIFAR10). To evaluate how the ensemble size affects the model performance, we modulate the ensembling frequency from *per epoch* to *10 epochs* and *100 epochs*, which results in an ensemble size of 100+ and 10+. As Figure 4.5 (left) shows, the smaller ensemble sizes lead to overall worse performance. This

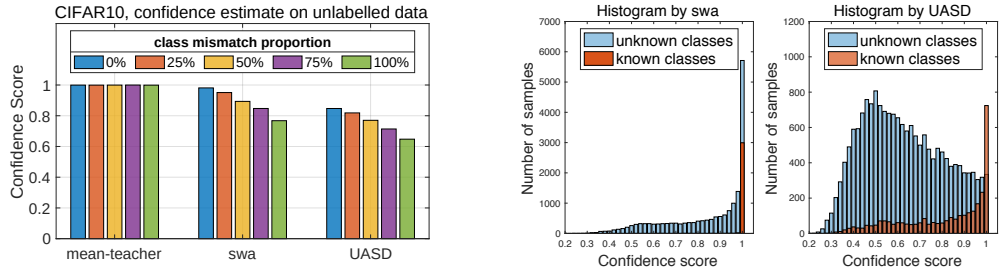



Figure 4.6: **Left**: Average confidence score on unlabelled data estimated by mean-teacher, SWA, UASD under varying mismatch proportion. **Right**: Histogram of confidence score by SWA, UASD.

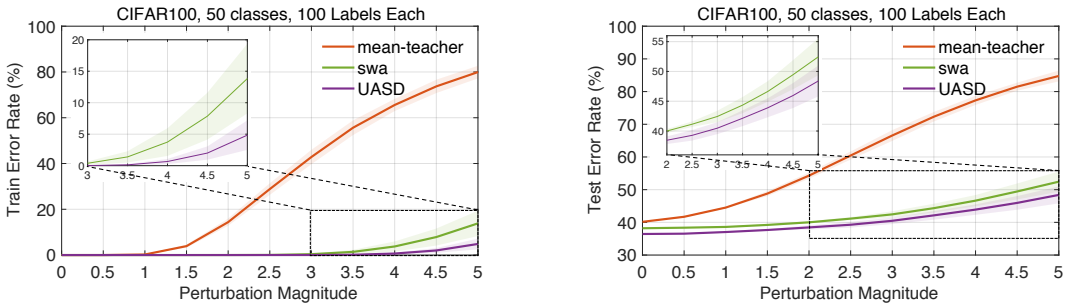



Figure 4.7: Model robustness under varying magnitude of perturbation on training data (**Left**) and test data (**Right**). Shaded area indicates the standard deviation over five randomly sampled perturbations.

suggests that ensemble size does matter, and aligns with our motivation of building a stronger ensemble model out of an increasing number of networks for deriving smoother and more reliable supervision signals.

**(II) Uncertainty-Aware Self-Distillation loss.** To evaluate how the loss formulation brings positive benefits, we conduct three ablative experiments: (1) w/o soft targets, which replaces soft targets with one-hot hard targets; (2) w/o OOD filter, which removes the OOD filter and takes all unlabelled data for training; (3) w/o soft + OOD, which uses one-hot hard targets and removes the OOD filter. Figure 4.5 (right) shows the ablative evaluation on CIFAR10, from which we analyse in two aspects as follows.

***(i) Effect of soft targets in Self-Distillation.*** When replacing the soft targets $q_t$ in Eq.(4.4) as hard targets, i.e. $\text{argmax}(q_t)$, we observe the ablative baseline (1) suffers large performance drops compared to the full model (4). How about learning from all unlabelled data with the soft targets? We find the performance still keeps in a reasonable range – see ablative baseline (2). This means that the soft targets yielded by UASD do provide rich information beyond the label supervision, which enables the network to learn from the unlabelled data in a self-supervised fashion. On one

side, rather than blindly fitting to overconfident class assignments, the network is encouraged to align with smoother predictive distributions, thus preserving the predictive uncertainty. On the other side, the soft targets serve to communicate the regularities of smoother class decision boundaries discovered by the preceding networks, therefore allowing to distill the knowledge from a committee of networks.

***(ii) Effect of OOD filter.*** When removing OOD filter, we observe the ablative baseline (2) suffer consistent performance drops compared to the full model (4). When removing the soft targets and OOD filter concurrently, we observe the worst performance drops – see ablative baseline (3). This indicates that discarding irrelevant unlabelled samples is important, and is especially vital when the unsupervised teaching signals are prone to overconfident, e.g. when using hard targets.

### 4.4.4 Further Analysis

To further understand why UASD is effective in SSL under class distribution mismatch. We compare three most competitive SSL methods for a more in-depth analysis, namely, (1) mean-teacher, (2) SWA and (3) UASD. We analyse in two different aspects as below.

**(I) Confidence calibration.** We compare the average confidence score (i.e. *maximum probability*) on the unlabelled data, estimated by the teaching signals in the end of training. In Figure 4.6 (left), it is evident that teaching signals given by mean-teacher are most *overconfident* – with the same level of *high confidence scores* under varying class distribution mismatch proportion. In contrast, confidence scores estimated by SWA, UASD are stratified to reflect uncertainty of the underlying class distribution. We further compare the histogram of confidence scores by SWA and UASD in Figure 4.6 (right). It shows UASD can better delimit between data from known and unknown classes. This indicates UASD yields *softer targets* (less overconfident), which are essential to guarantee robust SSL under class distribution mismatch. More analysis of confidence estimate is given in Figure 4.8.

**(II) Model generalisation.** To evaluate the model generalisation, we quantify the model robustness as the shifts of training and test error rates by adding perturbations on the networks. This is based on a well-known finding that convergence to a *wider optimum* typically leads to better *model generalisation* [Keskar et al., 2017; Chaudhari et al., 2017], while the width of optima can be approximately reflected as the model robustness under small perturbation [Izmailov et al., 2018]: $\theta(k,p) = \theta + k \cdot p$, where $p$ is the perturbation added on the model $\theta$ – a direction vector

with unit length drawn from a uniform distribution. We vary the scaling factor $k$ between $[0,5]$ to control the magnitude of perturbation. Figure 4.7 shows the model robustness against perturbations lie in an order as UASD>SWA>mean-teacher on both training and test data. This suggests UASD finds the widest optimum among the three and thus gives better model generalisability.

## 4.5 Summary

In this work, we systematically studied the more realistic semi-supervised learning (SSL) under class distribution mismatch, which poses a new challenge of how to maximise the value of unconstrained unlabelled data. To address this challenge, we proposed *Uncertainty-Aware Self-Distillation* (UASD), a novel SSL algorithm that utilises an on-the-fly accumulative ensemble to produce *soft targets* for joint *Self-Distillation* and *OOD filtering*. UASD consistently outperforms *six* state-of-the-art SSL methods on *three* image classification datasets. Although UASD has shown effectiveness in SSL under class distribution mismatch and suggests great value in practical use, we consider there are still several potential research directions to be further explored for addressing this new challenge: (1) tackle the class imbalance induced by class distribution mismatch; and (2) integrate class-incremental learning to cope with the unknown classes. Overall, our new problem setting along with the proposed approach open up many avenues for future research in SSL.

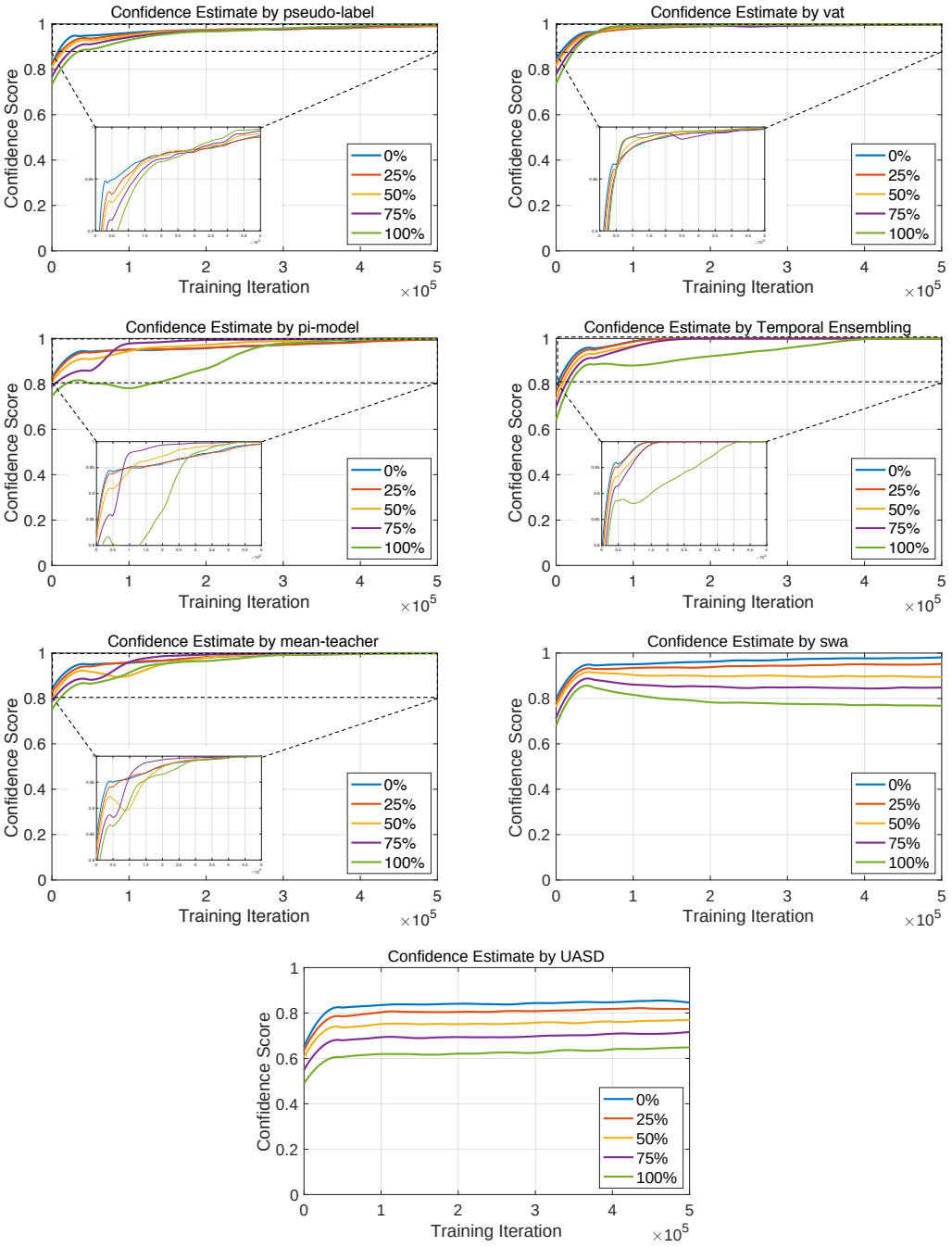



Figure 4.8: Average confidence score (i.e. maximum class probability) on unlabelled data, estimated by different teaching signals *during training* under varying class distribution mismatch proportion (i.e. 0, 25, 50, 75, 100%) on CIFAR10. Most SSL methods are prone to produce ***overconfident*** teaching signals, regardless of the underlying unlabelled class distribution. This hinders the possibility to be aware of uncertainty, and blindly reinforces the overconfident wrong class assignments on those irrelevant unlabelled samples. In contrast, UASD produces soft teaching signals that encode higher uncertainty, and exhibits different levels of confidence score that are clearly stratified to reflect the underlying class distribution mismatch proportions.

| | |
|---|---|
| **shared** | |
| training iterations | 500,000 |
| coefficient rampup from 0 util | 200,000 |
| learning decay factor | 0.2 |
| learning decay at iteration | 400,000 |
| **supervised baseline** | |
| initial learning rate | 0.003 |
| **pseudo-label** | |
| initial learning rate | 0.003 |
| max consistency coefficient | 1 |
| pseudo-label threshold | 0.95 |
| **VAT** | |
| initial learning rate | 0.003 |
| max consistency coefficient | 0.3 |
| VAT $\varepsilon, \xi$ | 6.0, $10^{-6}$ |
| **Π-Model** | |
| initial learning rate | 0.0003 |
| **Mean-Teacher** | |
| initial learning rate | 0.0004 |
| max consistency coefficient | 8 |
| Exponential moving average decay | 0.95 |
| **SWA** | |
| initial learning rate | 0.001 |
| max consistency coefficient | 8 |
| weight averaging interval | 5,000 |
| **UASD** | |
| initial learning rate | 0.001 |
| max distillation coefficient | 1 |

Table 4.2: Hyperparameter settings, which are inherited from the implementation by Oliver et al. Oliver et al. [2018], where hyperparameters are tuned on a validation set. An Adam optimiser is deployed with the same learning rate decay schedule. The following hyperparameters are used for all experiments. Note: we adopt the *same* ramp-up function for all methods.

| dataset | gaussian noise $\sigma = 0.15$ | horizontal flip $p = 0.5$ | random translation $[-2, +2]$ |
|---|---|---|---|
| CIFAR10 | ✓ | ✓ | ✓ |
| CIFAR100 | × | ✓ | ✓ |
| TinyImageNet | × | ✓ | ✓ |

Table 4.3: Data augmentation. Note: ZCA image pre-processing is *only* applied on CIFAR10.

| Dataset | $p\%$ | $K$ | $L_{\text{num}}$ | Labelled classes | Unlabelled classes |
|---|---|---|---|---|---|
| CIFAR10 | 0 | 400 | 6 | 2,3,4,5,6,7 | 4,5,6,7 |
| | 25 | | | | 0,5,6,7 |
| | 50 | | | | 0,1,6,7 |
| | 75 | | | | 0,1,8,7 |
| | 100 | | | | 0,1,8,9 |
| CIFAR100 | 50 | 50 | 100 | 0-50 | 25-75 |
| TinyImageNet | 50 | 100 | 100 | 0-100 | 50-150 |
| CIFAR100 + TinyImageNet | 86.5 | 100 | 100 | CIFAR100 | TinyImageNet |

Table 4.4: Evaluation protocols of SSL under class mismatch. $p\%$: Class distribution mismatch proportion among unlabelled data. $K$: number of known classes in labelled data. $L_{\text{num}}$: Labels per class.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | $24.03 \pm 0.75$ | | |
| pseudo-label | **22.37 ± 0.66** | **24.02 ± 0.86** | $25.37 \pm 0.86$ | $26.11 \pm 1.91$ | $26.29 \pm 0.71$ |
| VAT | **20.63 ± 1.77** | **23.08 ± 0.49** | **23.78 ± 0.70** | $25.52 \pm 0.84$ | $26.23 \pm 0.37$ |
| Π-Model | **21.56 ± 1.29** | $24.80 \pm 1.32$ | $25.92 \pm 1.61$ | $26.43 \pm 0.81$ | $26.61 \pm 0.79$ |
| Temporal Ensembling | **21.93 ± 0.43** | $24.23 \pm 0.96$ | $25.66 \pm 1.21$ | $26.33 \pm 0.56$ | $27.00 \pm 1.39$ |
| Mean-Teacher | **21.68 ± 0.88** | $24.13 \pm 1.22$ | $24.79 \pm 1.53$ | $25.90 \pm 1.00$ | $26.78 \pm 0.38$ |
| SWA | **21.63 ± 0.38** | **23.31 ± 0.85** | **23.70 ± 0.61** | **23.90 ± 0.85** | $24.11 \pm 0.65$ |
| UASD (ours) | **20.59 ± 0.51** | **21.34 ± 0.52** | **21.88 ± 0.69** | **22.39 ± 0.48** | **22.49 ± 0.90** |

Table 4.5: Evaluation under varying class distribution mismatch proportion on CIFAR10. Test error rates are reported at the point of lowest validation error. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**. Statistical significant tests of these methods as compared to the baseline are given in Table 4.6.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| pseudo-label | 0.0625 | 1.0000 | 0.1250 | 0.0625 | 0.0625 |
| VAT | 0.0625 | 0.1250 | 0.0625 | 0.0625 | 0.0625 |
| Π-Model | 0.0625 | 0.2500 | 0.0625 | 0.0625 | 0.0625 |
| Temporal Ensembling | 0.0625 | 0.8125 | 0.0625 | 0.0625 | 0.0625 |
| Mean-Teacher | 0.0625 | 1.0000 | 0.3125 | 0.0625 | 0.0625 |
| SWA | 0.0625 | 0.0625 | 0.3125 | 0.0625 | 0.8125 |
| UASD (ours) | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |

Table 4.6: Statistical significant tests under varying class distribution mismatch proportion on CIFAR10, which compare the results of each method to the results of the baseline. Note: results are the error rates at the point of lowest validation error.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | $23.82 \pm 0.61$ | | |
| pseudo-label | $\mathbf{22.70 \pm 0.42}$ | $24.42 \pm 0.87$ | $26.47 \pm 1.01$ | $27.82 \pm 1.10$ | $28.07 \pm 1.01$ |
| VAT | $\mathbf{23.07 \pm 0.49}$ | $27.27 \pm 1.36$ | $27.45 \pm 2.17$ | $28.46 \pm 2.62$ | $28.79 \pm 1.11$ |
| Π-Model | $\mathbf{22.97 \pm 0.46}$ | $26.48 \pm 0.66$ | $29.01 \pm 2.67$ | $28.19 \pm 0.97$ | $29.43 \pm 1.88$ |
| Temporal Ensembling | $\mathbf{22.45 \pm 0.59}$ | $25.33 \pm 0.81$ | $26.94 \pm 0.57$ | $27.59 \pm 0.62$ | $28.16 \pm 0.70$ |
| Mean-Teacher | $\mathbf{22.09 \pm 0.57}$ | $25.40 \pm 0.41$ | $26.46 \pm 0.78$ | $27.83 \pm 1.43$ | $29.09 \pm 1.44$ |
| SWA | $\mathbf{21.70 \pm 0.34}$ | $\mathbf{23.36 \pm 0.74}$ | $23.83 \pm 0.61$ | $24.15 \pm 0.90$ | $24.31 \pm 0.55$ |
| UASD (ours) | $\mathbf{20.55 \pm 0.41}$ | $\mathbf{21.66 \pm 0.71}$ | $\mathbf{22.01 \pm 0.78}$ | $\mathbf{22.31 \pm 0.65}$ | $\mathbf{22.63 \pm 0.78}$ |

Table 4.7: Evaluation under varying class distribution mismatch proportion on CIFAR10. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**. Statistical significant tests of these methods as compared to the baseline are given in Table 4.8.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| pseudo-label | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |
| VAT | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |
| Π-Model | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |
| Temporal Ensembling | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |
| Mean-Teacher | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |
| SWA | 0.0625 | 0.1250 | 1.0000 | 0.6250 | 0.0625 |
| UASD (ours) | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 |

Table 4.8: Statistical significant tests under varying class distribution mismatch proportion on CIFAR10, which compare the results of each method to the results of the baseline. Note: results are the error rates as the median of last 20 epochs.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | $23.82 \pm 0.61$ | | |
| size 10+ | **21.12 ± 0.69** | **22.64 ± 0.77** | **23.41 ± 0.56** | $24.09 \pm 0.72$ | $24.68 \pm 0.70$ |
| size 100+ | **20.57 ± 0.39** | **21.40 ± 0.41** | **22.59 ± 0.49** | **22.44 ± 0.70** | **22.80 ± 0.55** |
| size 1000+ (ours) | **20.55 ± 0.41** | **21.66 ± 0.71** | **22.01 ± 0.78** | **22.31 ± 0.65** | **22.63 ± 0.78** |

Table 4.9: Evaluation under varying class distribution mismatch proportion on CIFAR10. "size": ensemble size. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | $23.82 \pm 0.61$ | | |
| w/o both | **23.50 ± 0.86** | $24.78 \pm 0.64$ | $25.78 \pm 0.57$ | $26.05 \pm 0.79$ | $27.43 \pm 0.74$ |
| w/o soft | **21.84 ± 0.53** | **23.27 ± 0.60** | $24.67 \pm 0.60$ | $24.67 \pm 0.82$ | $25.52 \pm 0.92$ |
| w/o OOD | **21.19 ± 0.31** | **22.34 ± 0.52** | **22.61 ± 0.99** | **22.68 ± 0.56** | **23.11 ± 0.70** |
| Full UASD (ours) | **20.55 ± 0.41** | **21.66 ± 0.71** | **22.01 ± 0.78** | **22.31 ± 0.65** | **22.63 ± 0.78** |

Table 4.10: Evaluation under varying class distribution mismatch proportion on CIFAR10. "w/o both": w/o soft distillation and w/o OOD filter. "w/o soft": w/o soft distillation. "w/o OOD": w/o OOD filter. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

# Chapter 5

# Open-Set Cross-Domain Learning by Instance-Guided Context Rendering

---

Person re-identification (re-id) is a task of re-identifying a query person-of-interest, across non-overlapping cameras distributed over wide surveillance spaces [Gong et al., 2014]. Since the surge of deep representation learning, great boosts of re-id performance have been witnessed in an idealistic closed-world supervised learning testbed [Zheng et al., b; Xiao et al., 2016; Wang et al., 2016a; Zheng et al., c; Hermans et al., 2017; Chen et al., 2017c; Li et al., 2018; Sun et al., b; Chen et al.]: The rank-1 matching rate has reached 93.3% [Chen et al.] on the Market1501 benchmark [Zheng et al., b], as compared to 44.4% in 2015. However, this success relies heavily on an *unrealistic* assumption that the training and test data have to be drawn from the same camera network, i.e. the same domain. When deploying such re-id models to new domains, their performances often degrade significantly, mainly due to the inevitable domain gaps between datasets collected from independent surveillance camera networks. This weakness greatly restricts the generalisability of these *domain-specific* learning methods in real-world deployment, when manually labelling new identity population becomes prohibitively expensive at large scale. It is therefore essential to automate the domain-adaptive learnability with more advanced and robust *domain-generic* learning models.

The aforementioned problem, known as cross-domain person re-id, is gaining increasing attention [Peng et al., 2016; Wang et al., 2015; Ma et al., 2015; Wei et al., 2018; Deng et al., 2018; Wang et al., 2018b; Bak et al., 2018; Zhong et al.; Lin et al., a]. It raises a more challenging

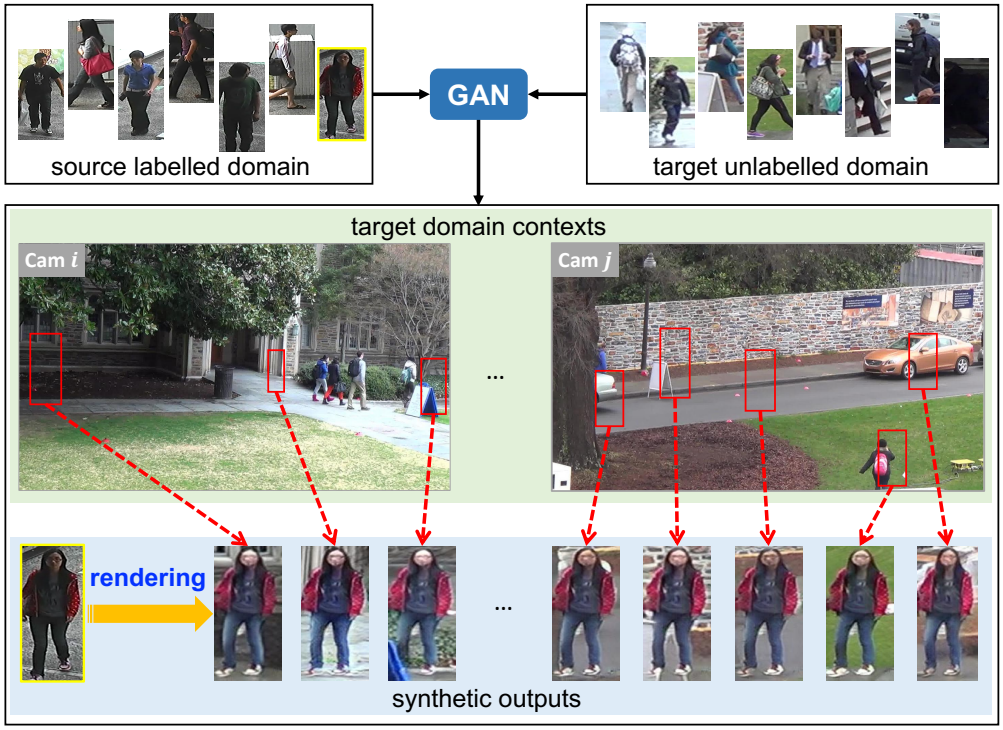



Figure 5.1: Motivation illustration. In open surveillance spaces, the *contextual variations* can be quite diverse, due to *wide-of-the-field imagery* and *varying times of the day*. Our approach learns to hallucinate the same persons in such diverse surveillance contexts, as if they were captured from different places and times in the target domain.

*open-set* unsupervised domain adaptation problem [Busto and Gall; Saito et al.], which requires to bridge the domain gap between two *disjoint* identity class spaces. The goal is to learn from source domain labelled data and target domain unlabelled data synergistically, so as to build more generalisable re-id model in the test target domain.

The state-of-the-art methods [Wei et al., 2018; Deng et al., 2018; Wang et al., 2018b; Zhong et al.; Lin et al., a; Bak et al., 2018; Liang et al.] can be categorised into three learning paradigms: (1) *feature-level distribution alignment*; (2) *image-level style transfer*; and (3) *hybrid image-level and feature-level learning*. The first paradigm [Wang et al., 2018b; Lin et al., a] generally seeks a common feature space for source-target distribution alignment with discriminative learning constraints. The second paradigm [Wei et al., 2018; Deng et al., 2018; Bak et al., 2018; Liang et al.] reduces the domain gap by employing GAN frameworks to transfer source images into target domain styles in a holistic manner. The last paradigm [Zhong et al.] unifies the complementary benefits of synthetic images by GAN and feature discriminative constraints in CNN. However, these existing paradigms all neglect to exploit the *rich contextual variations* as a potential domain bridge. In this work, we aim to utilise the contextual information for more effective

re-id model learning. This is motivated by our observation of complex environmental dynamics commonly existed in open public scenes (see Figure 5.1) – domain contexts are indeed quite diverse in surveillance spaces, given that the viewing conditions vary dramatically both *within* and *across* camera views, subjected to camera characteristics, wide-field-of-view imagery, and varying times of the day. We identify that the common weakness of existing GAN-based re-id methods lies in the insufficient data diversity – either *one* or a *pre-fixed* number of domain styles are captured in the final outputs. This is mainly caused by the *mode collapse* problem in GAN: merely a limited modes are plausibly generated. Our key idea is to rectify the aforementioned issue of *mode collapse* by rendering the source persons into diverse domain contexts, such that a large-scale *context augmented synthetic dataset* can be generated to train a re-id model in a supervised manner, without labelling any target domain data.

Specifically, we propose a novel Instance-Guided Context Rendering scheme, which augments the same source identity population with rich contextual variations reflected in the target domain. Our approach is unique in several perspectives. *First*, it effectively exploits *abundant unlabelled target instances* as guidance to render the source persons into different target domain contexts. This essentially captures the image-level domain drift in a more comprehensive way. *Second*, rather than optimising two-way mappings heavily with cycle consistency, we learn a simple *one-way mapping* through informative supervision signals. *Third*, compared to previous GAN-based re-id methods [Wei et al., 2018; Deng et al., 2018], our proposed *dual conditional* formulation naturally avoids *mode collapse* [Bansal et al.] to limited styles, and enables more diverse outputs. It transfers the same person into more realistic, finer-grained, and richer viewing conditions. The contextually more diverse synthetic imagery are ultimately utilised for re-id model learning to enhance visual invariance towards contextual variations in the target domain.

In summary, our **contribution** is two-fold:

- We propose a novel Instance-Guided Context Rendering scheme. To our best knowledge, it is *the first attempt in re-id* to tackle the image-level domain drift by injecting *rich contextual information* into the image generation process. It effectively augments the same source person images with diverse target domain contexts to construct a large-scale synthetic training set for re-id model learning in the unlabelled target domain.

- We design a dual conditional generative adversarial network. It effectively exploits abundant unlabelled target instances as contextual guidance to produce more plausible data with

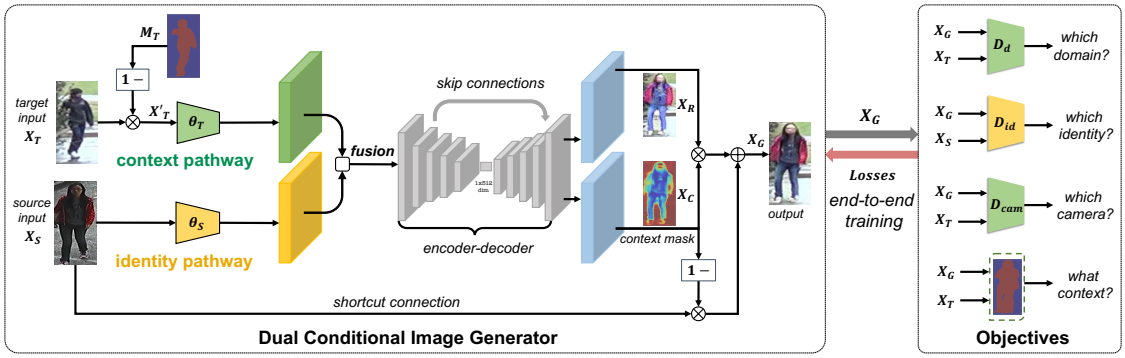



Figure 5.2: **Model overview.** We tackle the domain drift at the image level by learning to render the source person image $\boldsymbol{X}_S$ into diverse domain contexts explicitly guided by arbitrary target instances $\boldsymbol{X}_T$ sampled from the target domain (Section 5.1.2, Section 5.2).

richer *cross-* and *intra-domain contextual variations*. We conduct extensive experiments to validate our model design rationale, and show that our approach not only achieves competitive re-id performance on several re-id benchmarks in the cross-domain setting, but also generates photo-realistic person images with high fidelity and diversity.

## 5.1 Instance-Guided Context Rendering

We consider the problem of unsupervised domain adaptation in person re-id, which aims to adapt a re-id model learned from a labelled source dataset to an unlabelled target dataset. Our objective is to learn a generative mapping $G$ that reduces the domain discrepancy by rendering the same source person images into a diverse range of target domain contexts. As the final synthetic images are augmented with rich target contexts, a CNN model can simply be fine-tuned upon these data to enhance its generalisability in the unlabelled target domain.


Figure 5.3: **Deployment overview.** In deployment, the generator is applied to produce abundant images $\boldsymbol{X}_G$ for CNN training (Section 5.2.2).

### 5.1.1 Approach Overview.

Figure 5.2 illustrates our Instance-Guided Context Rendering scheme. Its main body is a dual conditional Generative Adversarial Network that takes in a pair of input images from two

domains for image generation (Section 5.1.2), and learns with informative supervision signals to render the source persons guided by different target instances (Section 5.2). We named our **C**ontext **R**endering Network as **CR-GAN** for short. For deployment (Figure 5.3), abundant data augmented with diverse context is exploited for re-id model learning in the pseudo target domain (Section 5.2.2).

### 5.1.2 Dual Conditional Image Generator

**Dual Conditional Mapping.** CR-GAN contains a dual conditional image generator that learns a one-way mapping to render the source images into desired target contexts by conditioning on **two** inputs: a source input $\boldsymbol{X}_S$ and a target input instance $\boldsymbol{X}_T$ to **guide** the context rendering effect. Formally, this dual conditional mapping is expressed as:

$$\boldsymbol{X}_G = G(\boldsymbol{X}_S, \boldsymbol{X}_T) \tag{5.1}$$

In essence, this dual conditional formulation is designed to fuse information flows from two domains, such that the same person in source input $\boldsymbol{X}_S$ can be rendered into the target context explicitly guided by the target instance $\boldsymbol{X}_T$. Overall, the whole mapping is built upon dual-path encoding and decoding, with a U-Net [Ronneberger et al., 2015] like encoder-decoder network in between, as detailed below.

**Dual-Path Encoding.** To enable instance-guided context rendering, we introduce an essential condition $\boldsymbol{X}_T$ to exploit abundant target instances as contextual guidance for generation. Concretely, we design a dual-path encoding structure to parameterise information flows from two domain separately (Figure 5.2) – **(1)** An *identity pathway* $\theta_S$ to encode source input $\boldsymbol{X}_S$; and **(2)** A *context pathway* $\theta_T$ to encode target input $\boldsymbol{X}_T$. Given that our aim is to exploit contextual information from target domain, we mask the target input $\boldsymbol{X}_T$ to retain mainly the background clutter. Specifically, we adopt the off-the-shelf human parsing model LIP-JPPNet [Gong et al.] to obtain a binary person mask, and apply spatial masking on $\boldsymbol{X}_T$ to filter out the target person:

$$\boldsymbol{X}'_T = \boldsymbol{X}_T \circ (\mathbf{1} - \boldsymbol{M}_T) \tag{5.2}$$

where $\circ$ is the Hadamard product; $\boldsymbol{M}_T$ is the person mask of input $\boldsymbol{X}_T$; $\boldsymbol{X}'_T$ contains mainly the background clutter. The pre-trained parsing network is kept frozen during training to provide a rough estimation on where the person is in the image. Since we do not have any groundtruth of the person masks, is also infeasible to update the parsing network in end-to-end training.

Through dual-path encoding, information flows from two domains are further fused by depth-wise concatenation: $[\theta_S(\boldsymbol{X}_S), \theta_T(\boldsymbol{X}'_T)]$, followed with an encoder-decoder network to selectively blend the visual information from two inputs. We construct the encoder-decoder network as a U-Net with skip connections to reuse latent representations from the encoder, which enforces the generator network to selectively preserve low-level visual structures from both conditional inputs. In particular, the foreground person in $\boldsymbol{X}_S$, the background clutter in $\boldsymbol{X}_T$ should both be picked by the generator as informative cues for image generation.

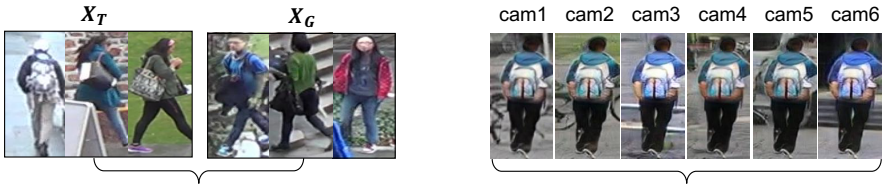
**Image Generation.** Given our aim to render the context in a region-dependent manner – keep the source person whilst augmenting background content, we employ a context mask to softly specify the region of contextual changes. Concretely, the generator outputs two parts: **(1)** A *residual map* $\boldsymbol{X}_R$ to model cross-domain discrepancy; and **(2)** A *context mask* $\boldsymbol{X}_C$ to modulate per-pixel intensity of context change, both of which are connected by a shortcut connection to reuse the source person in input $\boldsymbol{X}_S$. Such generic masking mechanisms are also adopted in recent literature, such as face animation [Pumarola et al.], motion manipulation [Zhao et al.]; while we particularly utilise the context mask to automatically learn the region selection of context rendering. The final generated output $\boldsymbol{X}_G$ is the sum of source input $\boldsymbol{X}_S$ and residual map $\boldsymbol{X}_R$ spatially weighted by the context mask $\boldsymbol{X}_C$:

$$\boldsymbol{X}_G = \boldsymbol{X}_R \circ \boldsymbol{X}_C + \boldsymbol{X}_S \circ (\mathbf{1} - \boldsymbol{X}_C) \tag{5.3}$$

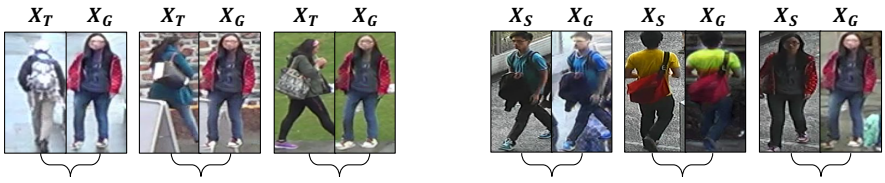
The generator is trained end-to-end to generate $\boldsymbol{X}_G$ with the same person identity as $\boldsymbol{X}_S$ in the context guided by $\boldsymbol{X}_T$.

## 5.2 Model Optimisation

The key idea of CR-GAN is to inject *context information* into image generation. This is motivated that context variations exist at multi-granularity – not only differ across domains, but also vary dramatically within and across different camera views. To learn such rich contexts, we impose *four* different supervision signals for optimisation, which work synergistically to learn (a) *cross-domain*, (b) *cross-camera*, and (c) *inner-camera context variations*, whilst (d) retaining the *source identity* – illustrated in Figure 5.4 and elaborated below.




(a) $\mathcal{L}_{\text{adv}}$ reduces *domain gap*. (b) $\mathcal{L}_{\text{cam}}$ learns *camera variations*.




(c) $\mathcal{L}_{\text{con}}$ learns *context variations*. (d) $\mathcal{L}_{\text{id}}$ retains *source identity*.

Figure 5.4: Schematic illustration of learning objectives.

### 5.2.1 Learning Objectives

**Adversarial Loss.** To mitigate the *cross-domain* contextual gap, the generator $G$ is trained against a domain discriminator $D_d$ in an adversarial minimax manner [Goodfellow et al.]:

$$\mathcal{L}_{\text{adv}} = \min_{G} \max_{D_d} \log D_d(\boldsymbol{X}_T) + \log(1 - D_d(G(\boldsymbol{X}_S, \boldsymbol{X}_T))) \tag{5.4}$$

where $\mathcal{L}_{\text{adv}}$ aligns the generated data distribution with the target data distribution globally to reduce the domain gap. That is, $\mathcal{L}_{\text{adv}}$ serves as an essential loss term that encourages the generator $G$ to generate person images similar to the target domain images. We adopt PatchGAN [Isola et al.] to discriminate stylistic statistics at the scale of patches, hence penalising the distribution discrepancy in a more structural way.

**Camera Loss.** To capture the *cross-camera* context variations induced by camera characteristics – e.g. *colour tones* – a camera loss is imposed to constrain the camera styles:

$$\mathcal{L}_{\text{cam}} = -\log(p(y_c|\boldsymbol{X}_G)) \tag{5.5}$$

where $y_c$ is camera label of $\boldsymbol{X}_T$ that can be accessible from the image metadata without manual annotations. $\mathcal{L}_{\text{cam}}$ is derived by a camera discriminator $D_{cam}$ trained to classify camera labels.

**Context Loss.** While capturing context variations across domains, cameras, the generator should also learn the *inner-camera* context changes with content details. Accordingly, we adopt masked reconstruction errors to constrain foreground, background similar to input $\boldsymbol{X}_S, \boldsymbol{X}_T$ respectively:

$$\mathcal{L}_{\text{con}} = ||(\boldsymbol{X}_G - \boldsymbol{X}_S) \circ \boldsymbol{M}_F||_2 + ||(\boldsymbol{X}_G - \boldsymbol{X}_T) \circ \boldsymbol{M}_B||_2 \tag{5.6}$$

where $\boldsymbol{M}_F, \boldsymbol{M}_B$ are the foreground, background person masks of $\boldsymbol{X}_S$ extracted by human parsing model. $\mathcal{L}_{\text{con}}$ particularly encourages to retain the source person, whilst augmenting more *diverse background clutters* explicitly guided by arbitrary target instance $\boldsymbol{X}_T$ from the target domain.

**Identity Loss.** As the source person identity in input $\boldsymbol{X}_S$ should be preserved in output $\boldsymbol{X}_G$, we impose an identity classification error to constrain the person identity in $\boldsymbol{X}_G$:

$$\mathcal{L}_{\text{id}} = -\log(p(y_j|\boldsymbol{X}_G)) \tag{5.7}$$

where $y_j$ is the identity label of $\boldsymbol{X}_S$; $\mathcal{L}_{\text{id}}$ is derived by an identity discriminator $D_{id}$ – a standard re-id CNN backbone trained to predict the source person identities.

**Overall Objective.** CR-GAN is trained with joint optimisation of four losses (Eq. (5.4),(5.5),(5.6) and (5.7)) for their complementary benefits in constraining the image generation:

$$\mathcal{L}_{GAN} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} \tag{5.8}$$

where $\lambda_{\text{adv}}, \lambda_{\text{id}}, \lambda_{\text{cam}}, \lambda_{\text{con}}$ are hyper-parameters to control the relative importance of each loss. We set $\lambda_{\text{id}}$=$\lambda_{\text{cam}}$=1, $\lambda_{\text{adv}}$=2, $\lambda_{\text{con}}$=5 to keep the losses in similar value range.

### 5.2.2 Model Training and Deployment

CR-GAN is optimised similar to the standard GAN models. For deployment, $D_{id}$ – a standard backbone ResNet50 [He et al., 2016] – is fine-tuned upon abundant context augmented data generated by CR-GAN (Figure 5.3). All synthetic data is randomly produced on-the-fly by feeding arbitrary source-target image pairs to CR-GAN, therefore eschewing the need of storing an extremely large-scale synthetic dataset. After fine-tuning, the backbone network $D_{id}$ is deployed to extract features for re-id matching in the target domain. The model training is summarised in Algorithm 3.

**Remark.** It is worth mentioning that a two-stage training is necessary in our context. If we integrate generative learning of image generation and discriminative learning of feature representation as one-stage end-to-end training, the discriminative re-id network would receive noisy and corrupted generated person images. The corrupted input images could harm the model and further lead to its failure in distinguishing different person identities.

**Algorithm 3** Instance-Guided Context Rendering.

**I. Initialisation:** Pre-train $D_{id}, D_{cam}$ with labels.

**II. Train the image generator $G$:**

**Input:** Source dataset $\mathcal{D}_S$, target dataset $\mathcal{D}_T$.

**Output:** An image generator $G$.

**for** $t = 1$ **to** *max_gan_iter* **do**

Feedforward mini-batch of input pairs $(\boldsymbol{X}_S, \boldsymbol{X}_T)$ to $G$.

Update $D_d$ (Eq. (5.4)) and update $G$ for $k$ times (Eq. (5.8)).

**end for**

**III. Fine-tune $D_{id}$ on synthetic data:**

**for** $t = 1$ **to** *max_cnn_iter* **do**

Random context rendering: $\boldsymbol{X}_G = G(\boldsymbol{X}_S, \boldsymbol{X}_T)$.

Update $D_{id}$ on $\boldsymbol{X}_G$ using identity label of $\boldsymbol{X}_S$.

**end for**

## 5.3 Discussion

As compared to existing **Unsupervised Domain Adaptation** (UDA) techniques that rely on either *feature-level adaptation* [Tzeng et al., 2014; Long et al., 2015; Sun and Saenko, 2016; Ganin et al., 2016; Tzeng et al., 2017; Xie et al., 2018] or *image-level adaptation* [Bousmalis et al., 2017; Shrivastava et al., 2017] to mitigate the cross-domain distribution discrepancy, our approach also learns to transform the image styles, but particularly focuses on enriching the diversity of synthetic images to facilitate more effective domain adaptation in re-id. While comparing to existing **Image-to-Image Translation** (I2I) techniques that mainly aims to transform images from original styles to new styles – as represented by Pix2Pix [Isola et al.], CycleGAN [Zhu et al.], StarGAN [Choi et al., 2018], MUNIT [Huang et al.] and DRIT [Lee et al., a], our CR-GAN is especially driven by the goal of diversifying the generated outputs to produce more synthetic training data for effective domain adaptation.

Overall, our CR-GAN has the following merits to benefit cross-domain re-id model learning: **(I)** Instead of controlling the rendering effects with a fixed set of category labels, such as camera labels, we leverage *abundant unlabelled instances* $\boldsymbol{X}_T$ from target domain as contextual guidance to inject contextual variations. This naturally avoids *mode collapse* to limited fixed

styles, and synthesises more diverse target domain contexts for learning a domain-generic re-id model. **(II)** Rather than changing the domain contexts holistically, our rendering effects are region-dependent. In particular, the *background clutter* is significantly modified with *structural change*; while the *foreground person* is slightly inpainted with *colour change* to capture the domain drift. Such rendering effects effectively retain the source identity, whilst augmenting richer contexts for re-id model learning in the synthetic pseudo target domain. **(III)** By fusing two inputs through dual-path encoding at the lower layers, the generator network is enforced to learn the selective preservation of low-level visual structures from both inputs, therefore enhancing the modelling capacity to produce synthetic training data in higher fidelity and diversity.

It is worth mentioning that our proposed approach can not only be applied for open-set cross-domain learning in person re-id, but can also generalise to other recognition tasks that require to tackle the domain gap across two domains with non-overlapping label spaces. As the source domain label information and the target domain context are merged in the generated synthetic data, a discriminative recognition model trained upon these data can become both discriminative to the task-relevant label information and invariant to the task-irrelevant domain variations.

## 5.4 Experiments

### 5.4.1 Experimental settings

**Implementation Details.** To train CR-GAN, we use the Adam solver [Kingma and Ba, 2014] with a mini-batch size of 32. The learning rate is set to 0.0002 in the first half of training and linearly decayed to 0 in the second half. To build up the image generator, Instance Normalisation (IN) [Ulyanov et al., 2017] is used in the U-Net decoder. IN is neither applied in two separate encoding pathways nor the U-Net encoder, which allows to retain the stylistic information before decoding. The two pathways for dual condition are parameterised as separate convolutional layers. To improve the training stability of GAN, we add one additional Gaussian noise layer as the input layer in the domain discriminator. We employ LSGAN [Mao et al.] as the GAN formulation and adopt the domain discriminator same as PatchGAN [Isola et al.] to discriminate at the scale of patches. To stabilise the training, the image generator is updated twice every iteration in the second half of training. We use the standard ImageNet [Deng et al., 2009] pre-trained ResNet50 as the identity discriminator $D_{id}$. The camera discriminator $D_{cam}$ is an extremely lightweight CNN classifier with 5 layers. More details on network architectures are given in Table 5.9, 5.10

and 5.11.

**Training Procedures.** As aforementioned in Algorithm 3, the training process is divided into three steps. First, for initialisation, we pre-train the identity discriminator (ResNet50), camera discriminator for 30,000 iterations. Second, we train the image generator, domain discriminator from scratch for 60,000 iterations. Third, we fine-tune the ResNet50 using synthetic data produced by the image generator on-the-fly. We only apply random flipping as data augmentation. After training, the ResNet50 is used as the backbone network to extract feature.

**Evaluation Metrics.** We adopt several metrics to comprehensively evaluate our model in two aspects: **(1)** To evaluate the re-id matching performance, we adopt the standard *Cumulative Match Characteristic* (**CMC**) and *mean Average Precision* (**mAP**) as evaluation metrics. We report results on *single-query* based on the ranking order of cross-camera pairwise matching distances computed based on features extracted from the re-id CNN model. **(2)** To measure the visual quality of synthesis, we adopt the following two evaluation metrics: **(i)** *LPIPS Distance* (**LPIPS**) [Zhang et al., 2018a] measures the *image translation diversity*, which is correlated with human perceptual similarity. We used the default ImageNet pre-trained AlexNet to extract feature in evaluation. **(ii)** *Fréchet Inception Distance* (**FID**) [Heusel et al.] measures the *image fidelity* by quantifying the distribution discrepancy between generated data and real data. We used the default ImageNet pre-trained Inception to extract feature in evaluation.

**Datasets.** We adopt three standard re-id benchmarks for evaluation (Figure 5.5). **(1) Market1501** [Zheng et al., b] contains 1,501 identities captured by 6 different cameras. The training set includes 751 identities and 12,936 images. The testing set includes 750 identities, with 3,368 images in the probe set and 19,732 images in the gallery set. **(2) DukeMTMCreID** [Ristani et al.; Zheng et al., c] contains 1,404 identities captured by 8 different cameras. The training set includes 702 identities and 16,522 images. The testing set includes 702 identities, with 2,228 images in the probe set and 17,661 images in the gallery set. **(3) CUHK03** [Li et al., d] contains 1,467 identities and 14,097 images in total. We use the auto-detected version.

### 5.4.2 Ablative Model Evaluation

To validate our model design rationale, we conduct ablative study on two different domain pairs: Market1501 $\rightarrow$ DukeMTMCreID, DukeMTMCreID $\rightarrow$ Market1501.

**Effect of Dual Condition.** Introducing abundant target instances as contextual guidance is the

(a) Market1501. (b) DukeMTMCreID. (c) CUHK03.

Figure 5.5: Example images from three re-id benchmarks.


Figure 5.6: Qualitative visual evaluation. Given source image $\boldsymbol{X}_S$, (a) baseline (w/o dual condition) *collapses* to uniform context, due to lack of *contextual guidance*; while (b) CR-GAN augments the same person with diverse contexts explicitly guided by target instances $\boldsymbol{X}_T$.

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics | LPIPS | FID | LPIPS | FID |
| Source-Target data | 0.458 | 0.330 | 0.458 | 0.330 |
| w/o dual condition | 0.196 | 0.065 | 0.210 | 0.137 |
| **CR-GAN** | **0.281** | **0.058** | **0.269** | **0.096** |

Table 5.1: Quantitative visual evaluation on image quality. **LPIPS**: *image perceptual similarity*, higher is better. **FID**: *distribution discrepancy*, lower is better. LPIPS / FID in "Source-Target data" represents the upper bound. Best results are in **bold**.

*key factor* that enables an Instance-Guided Context Rendering process. To validate this factor, we compare our dual conditional mapping (CR-GAN) with an ablative baseline that takes in merely the source input $\boldsymbol{X}_S$ (w/o dual condition). Figure 5.6 shows that: **(1)** Although the baseline transforms the context, all the generated images *collapse* to the same context; **(2)** CR-GAN, on the contrary, acts as a much stronger data generator to augment the same person with a more diverse range of domain contexts. This is in line with our visual quantitative results in Table 5.1, where CR-GAN obtains much higher LPIPS, i.e. more diverse outputs, compared to the baseline. This shows compellingly the benefit of our *dual conditional* formulation to exploit abundant

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics (%) | R1 | mAP | R1 | mAP |
| Direct Transfer | 36.9 | 20.5 | 47.5 | 20.0 |
| w/o dual cond | 43.3 | 24.8 | 55.5 | 27.0 |
| **CR-GAN** | **52.2** | **30.0** | **59.6** | **29.6** |
| w/o dual cond+LMP | 48.7 | 27.6 | 59.2 | 28.5 |
| **CR-GAN+LMP** | **56.0** | **33.3** | **64.5** | **33.2** |

Table 5.2: Ablation study of dual condition in re-id. "Direct Transfer": CNN trained with only labelled source data; "w/o dual cond": without dual condition; LMP: a pooling strategy [Deng et al., 2018] to reduce noisy signals induced by fake synthetic images at test time.

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics (%) | R1 | mAP | R1 | mAP |
| w/o identity loss | 31.9 | 15.4 | 32.8 | 11.8 |
| w/o camera loss | 48.8 | 28.6 | 53.6 | 26.0 |
| w/o context loss | 48.5 | 28.8 | 57.4 | 28.7 |
| **CR-GAN** | **52.2** | **30.0** | **59.6** | **29.6** |

Table 5.3: Ablation study on individual effect of each loss in re-id.

target instances as contextual guidance in the image generation.

To evaluate the benefit of context rendering effects in re-id, we compare CR-GAN with the ablative baseline. Table 5.2 shows that **(1)** Introducing our dual conditional formulation significantly boosts the re-id performance, with improved margins of 8.9% (52.2-43.3) / 4.1% (59.6-55.5) in R1 on DukeMTMCreID / Market1501. **(2)** The improvement remains in the use of LMP, with improved margins of 7.3% (56.0-48.7) / 5.3% (64.5-59.2) in R1. This indicates that re-id model learning with more contextual variations is indeed helpful to boost the cross-domain model robustness.

**Effect of Different Losses.** In addition to the standard adversarial loss, CR-GAN is trained with *three* different losses. To validate the necessity of using these losses in re-id, we conduct ablative comparison by eliminating individual loss from the overall objective. Table 5.3 shows that: **(1)** Removing any of the loss leads to undesired performance drop; **(2)** All losses work synergistically, with their joint optimisation to achieve the best performance. **(3)** These results are in line with our loss design rationale: All losses serve to exploit complementary information in model optimisation (Figure 5.4), and thus give their desired performance gains to generate


Figure 5.7: Qualitative visual evaluation. Given source image $\boldsymbol{X}_S$, (a) SPGAN [Deng et al., 2018] transforms the image into merely *one uniform style*; while (b) our CR-GAN renders the source persons into varying contexts: different *background clutters*, *colour tones* and *lighting conditions*.

better data for re-id model learning.

### 5.4.3 Analysis on GAN-based Methods

To isolate and analyse the pure effect of image-level domain adaptation in re-id, we compare our model with GAN-based methods for ablative analysis in this section.

**Qualitative Visual Analysis.** To understand how context information is brought to benefit the re-id model learning, we first visually compare the synthetic images produced by our CR-GAN with SPGAN [Deng et al., 2018]: a *representative* re-id method based upon *CycleGAN*. As Figure 5.7 shows, compared to merely one possible output given by SPGAN, CR-GAN can produce more diverse outputs. This informs that CR-GAN indeed serves as a much stronger *synthetic data generator* to augment much more contextual variations and thus produce a synthetic training set of much larger-scale. More qualitative results of CR-GAN on four different domain pairs as further shown in Figure 5.10, 5.11, 5.12, 5.13. Overall, our visualisation shows that CR-GAN is capable of producing abundant data augmented with different *background clutters*, *colour tones* and *lighting conditions*, explicitly guided by target instances randomly sampled from the target domain.

**Quantitative Visual Analysis.** To evaluate the visual quality quantitatively, we further compare CR-GAN with SPGAN based on the *synthetic data released by the authors*. Table 5.4 indicates that: **(1)** Both CR-GAN and SPGAN have lower and better FID compared to the FID between the source and target data. This informs that after style adaptation, the cross-domain distribution discrepancy is mitigated with both methods. **(2)** Compared to SPGAN, CR-GAN has much lower FID and higher LPIPS. This indicates CR-GAN can generate images of better fidelity and higher

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics | LPIPS | FID | LPIPS | FID |
| Source-Target data | 0.458 | 0.330 | 0.458 | 0.330 |
| SPGAN [Deng et al., 2018] | 0.099 | 0.171 | 0.099 | 0.115 |
| **CR-GAN** | **0.281** | **0.058** | **0.269** | **0.096** |

Table 5.4: Quantitative visual evaluation on image quality. **LPIPS**: *image perceptual similarity*, higher is better. **FID**: *distribution discrepancy*, lower is better. Best results are in **bold**.

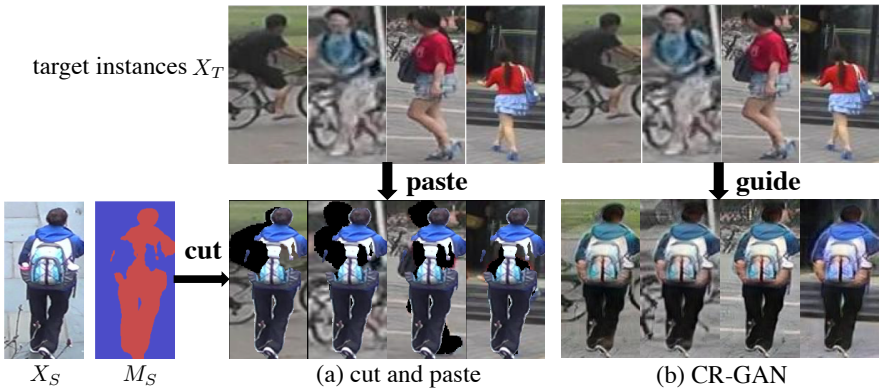



Figure 5.8: Synthetic images by (a) "*cut and paste*" and (b) our CR-GAN. $X_S$: source image; $X_T$: target image; $M_S$: parsing mask of $X_S$.

diversity.

**Analysis on Re-id Matching.** To further justify how our synthetic contextual variations benefit cross-domain re-id learning, we compare CR-GAN with three state-of-the-art GAN-based re-id methods: PTGAN [Wei et al., 2018], SPGAN [Deng et al., 2018], M2M-GAN [Liang et al.] on two domain pairs. All these models are trained on the *same source datasets* under the *same learning paradigm*: a GAN – an image generator – is first trained to synthesise images, a CNN is then fine-tuned upon the synthetic data for model adaptation. Table 5.6 shows that CR-GAN achieves the best cross-domain re-id performance. It is worth pointing out that previous methods generally collapse to fixed style(s): one homogenous domain style (PTGAN, SPGAN), or a pre-defined set of camera styles (M2M-GAN). In contrast, CR-GAN augments much rich contextual variations for re-id model learning.

### 5.4.4 Analysis on Image Synthesis Methods

We additionally illustrate the superiority of using CR-GAN to produce realistic synthetic data in comparison to an easy "*cut, paste and learn*" [Dwibedi et al.], which is an image synthesis

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics (%) | R1 | mAP | R1 | mAP |
| Direct Transfer | 36.9 | 20.5 | 47.5 | 20.0 |
| cut, paste and learn [Dwibedi et al.] | 21.6 ↓ | 9.0 ↓ | 26.5 ↓ | 11.3 ↓ |
| **CR-GAN** | **52.2** | **30.0** | **59.6** | **29.6** |
| **CR-GAN+LMP** | **56.0** | **33.3** | **64.5** | **33.2** |

Table 5.5: Ablation study in comparison to "*cut, paste and learn*".

| S → T | Market→Duke | | Duke→Market | |
|---|---|---|---|---|
| Metrics (%) | R1 | mAP | R1 | mAP |
| PTGAN [Wei et al., 2018] | 27.4 | - | 38.6 | 66.1 |
| SPGAN [Deng et al., 2018] | 41.1 | 22.3 | 51.5 | 22.8 |
| M2M-GAN [Liang et al.] | 49.6 | 26.1 | 57.5 | 26.8 |
| **CR-GAN** | **52.2** | **30.0** | **59.6** | **29.6** |
| SPGAN+LMP [Deng et al., 2018] | 46.4 | 26.2 | 57.7 | 26.7 |
| M2M-GAN+LMP [Liang et al.] | 54.4 | 31.6 | 63.1 | 30.9 |
| **CR-GAN+LMP** | **56.0** | **33.3** | **64.5** | **33.2** |

Table 5.6: Evaluation on GAN-based methods in the cross-domain re-id settings. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue.

approach originally proposed for instance detection. Specifically, we first *cut* the source person segment and *paste* it to the target background. Then, we train the re-id model upon the "*cut and paste*" synthetic data. Figure 5.8 illustrates that the "*cut and paste*" synthetic data not only contains various artifacts – some identity relevant cue (e.g. backpack) is missing due to incomplete person mask; but it also cannot capture the lighting nor colour tones of the target domain. These limitations are in line with its weaker performance as shown in Table 5.5, where "*cut, paste and learn*" yields even worse re-id results than "Direct Transfer". Overall, this demonstrates the necessity of designing our CR-GAN to generate synthetic training data of higher fidelity and diversity for enhancing the cross-domain generalisability.

### 5.4.5 Comparison with the State-of-the-art

**Competitors.** We compare our CR-GAN with 12 state-of-the-art methods. To ensure a ***like-to-like fair comparison***, we compare these methods by categorising them into four groups:

| Types | Source → Target | Market1501 → DukeMTMCreID | | | | DukeMTMCreID → Market1501 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Metrics (%) | R1 | R5 | R10 | mAP | R1 | R5 | R10 | mAP |
| Shallow | LOMO | 12.3 | 21.3 | 26.6 | 4.8 | 27.2 | 41.6 | 49.1 | 8.0 |
| | BoW | 17.1 | 28.8 | 34.9 | 8.3 | 35.8 | 52.4 | 60.3 | 14.8 |
| | UMDL | 18.5 | 31.4 | 37.6 | 7.3 | 34.5 | 52.6 | 59.6 | 12.4 |
| Image | PTGAN | 27.4 | - | 50.7 | - | 38.6 | - | 66.1 | - |
| | SPGAN+LMP | 46.4 | 62.3 | 68.0 | 26.2 | 57.7 | 75.8 | 82.4 | 26.7 |
| | M2M-GAN+LMP | 54.4 | - | - | 31.6 | 63.1 | - | - | 30.9 |
| | **CR-GAN+LMP** | **56.0** | **70.5** | **74.6** | **33.3** | **64.5** | **79.8** | **85.0** | **33.2** |
| Feature | PUL* | 30.0 | 43.4 | 48.5 | 16.4 | 45.5 | 60.7 | 66.7 | 20.5 |
| | TJ-AIDL† | 44.3 | 59.6 | 65.0 | 23.0 | 58.2 | 74.8 | 81.1 | 26.5 |
| | MMFA† | 45.3 | 59.8 | 66.3 | 24.7 | 56.7 | 75.0 | 81.8 | 27.4 |
| | BUC* | 47.4 | 62.6 | 68.4 | 27.5 | **66.2** | 79.6 | 84.5 | 38.3 |
| | TAUDL* | **61.7** | - | - | **43.5** | 63.7 | - | - | **41.2** |
| Hybrid | HHL | 46.9 | 61.0 | 66.7 | 27.2 | 62.2 | 78.8 | 84.0 | 31.4 |
| | SPGAN+TAUDL | 66.1 | 80.0 | 83.2 | 47.2 | 66.5 | 81.8 | 86.6 | 38.5 |
| | **CR-GAN+TAUDL** | **68.9** | **80.2** | **84.7** | **48.6** | **77.7** | **89.7** | **92.7** | **54.0** |

Table 5.7: Evaluation on Market1501, DukeMTMCreID in comparison to the state-of-the-art unsupervised cross-domain re-id methods. *: Not use auxiliary source training data. †: Use auxiliary source attribute labels for training. "-": no reported results. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue. Note that HHL uses StarGAN [Choi et al., 2018] to generate synthetic training images.

| Types | Source → Target | CUHK03 → Market1501 | | | | CUHK03 → DukeMTMCreID | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Metrics (%) | R1 | R5 | R10 | mAP | R1 | R5 | R10 | mAP |
| Image | PTGAN | 31.5 | - | 60.2 | - | 17.6 | - | 38.5 | - |
| | SPGAN | 42.3 | - | - | 19.0 | - | - | - | - |
| | **CR-GAN** | **58.5** | **75.8** | **81.9** | **30.4** | **46.5** | **61.6** | **67.0** | **26.9** |
| Feature | TAUDL* | **63.7** | - | - | **41.2** | **61.7** | - | - | **43.5** |
| Hybrid | HHL | 56.8 | 74.7 | 81.4 | 29.8 | 42.7 | 57.5 | 64.2 | 23.4 |
| | **CR-GAN+TAUDL** | **78.3** | **89.4** | **93.0** | **56.0** | **67.7** | **79.4** | **83.4** | **47.7** |

Table 5.8: Evaluation on CUHK03 to Market1501 / DukeMTMCreID adaption compared to state-of-the-art unsupervised cross-domain re-id methods. *: Not use source data. "-": no reported results. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue.

**(a)** *shallow methods using hand-crafted features*: LOMO [Liao et al.], BoW [Zheng et al., b], UMDL [Peng et al., 2016]; **(b)** *image-level learning methods*: PTGAN [Wei et al., 2018],

SPGAN [Deng et al., 2018], M2M-GAN [Liang et al.], which use GANs for style adaptation; **(c)** *feature-level learning methods*: PUL [Fan et al., 2018], TJ-AIDL [Wang et al., 2018b] MMFA [Lin et al., a], BUC [Lin et al., b], TAUDL [Li et al., b], which use additional discriminative constraints in CNN; **(d)** *hybrid learning methods*: HHL [Zhong et al.], which combine the benefits of group **(b)** and **(c)**.

It is worth noting that the learning paradigms in group **(b)**, **(c)** are essentially *orthogonal*: learning is performed either in *image space* or *feature space*. Therefore, these two paradigms should be complementary when unified in a hybrid formulation. To testify the generalisability of CR-GAN in a hybrid formulation, we add an additional comparison by unifying CR-GAN / SPGAN with the best performer TAUDL in group **(c)**. We first train the CNN with synthetic data generated by CR-GAN / SPGAN, then apply TAUDL with the pre-trained CNN in the target domain. Such hybrid formulations are denoted as CR-GAN+TAUDL / SPGAN+TAUDL, respectively.

**Evaluation on Market1501 / DukeMTMCreID.** Table 5.7 shows comparative results on two domain pairs. It can be observed that **(1)** CR-GAN performs best in the *image-level* learning paradigm; **(2)** When deploying CR-GAN in a hybrid formulation (CR-GAN+TAUDL), we earn the best re-id performance due to complementary benefits of two learning paradigms. In particular, CR-GAN+TAUDL boosts the performance over TAUDL with margins of 7.2% (68.9-61.7) / 14.0% (77.7-63.7) in R1 on DukeMTMCreID / Market1501. These results not only indicate the benefit of unifying ***GAN-based*** *image-level learning* and ***CNN-based*** *feature-level learning* into unsupervised cross-domain re-id, but more importantly justify our rationale of augmenting richer contextual variations to enable learning a more effective re-id model in the applied domain.

**Evaluation on CUHK03 to Market1501 / DukeMTMCreID.** Table 5.8 shows comparative results on model adaptation from CUHK03, where there exists larger domain gaps between the source and target domains (Figure 5.5). It can be seen that **(1)** CR-GAN clearly outperforms the best image-level competitor SPGAN with large margins; **(2)** When deploying in a hybrid formulation, CR-GAN+TAUDL outperforms the best hybrid competitor HHL with large margins of 21.5% (78.3-56.8), 25.0% (67.7-42.7) in R1 on Market1501 / DukeMTMCreID respectively. These collectively suggest the significant advantages of exploiting the synthetic data by CR-GAN in cross-domain re-id model learning.

## 5.5 Summary

We presented a novel Instance-Guided Context Rendering scheme for cross-domain re-id model learning. Through a carefully-designed dual conditional mapping, abundant target instances are exploited as contextual guidance for image generation. We conducted extensive ablative analysis to validate our model design rationale, and show the best performance over existing GAN-based re-id methods. Our like-to-like comparison with the state-of-the-art methods demonstrates the great advantage of our model when flexibly deploying in a hybrid systematic formulation. Overall, CR-GAN serves as a generic generator to augment abundant domain contexts for re-id model learning in practice.

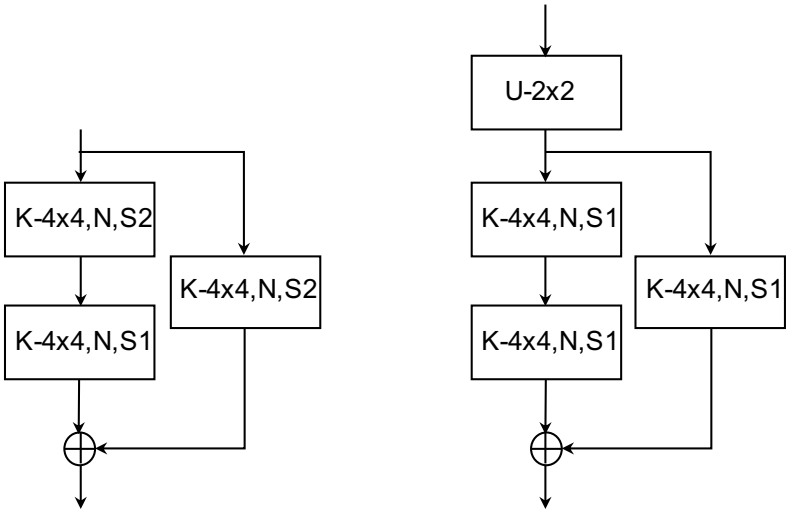



Figure 5.9: Left: Downsampling residual block. Right: Upsampling residual block. Note: conv layer is introduced in the shortcut connection as the number of feature maps in input and output are not necessarily the same in the U-Net.

| **Part Name** | **Input → Output Shape** | **Layer Description** |
|---|---|---|
| Encoding | $(H,W,3) \rightarrow (\frac{H}{2},\frac{H}{2},64)$ | context pathway: conv-(K-4×4, N-64, S-2, P-0, PS, LReLU) |
| | $(H,W,3) \rightarrow (\frac{H}{2},\frac{H}{2},64)$ | identity pathway: conv-(K-4×4, N-64, S-2, P-0, PS, LReLU) |
| U-Net (encoder) | $(\frac{H}{2},\frac{H}{2},128) \rightarrow (\frac{H}{4},\frac{W}{4},128)$ | res-(K-4×4, N-128, P-0, PS, LReLU) |
| | $(\frac{H}{4},\frac{H}{4},128) \rightarrow (\frac{H}{8},\frac{W}{8},256)$ | res-(K-4×4, N-256, P-0, PS, LReLU) |
| | $(\frac{H}{8},\frac{H}{8},256) \rightarrow (\frac{H}{16},\frac{W}{16},512)$ | res-(K-4×4, N-512, P-0, PS, LReLU) |
| | $(\frac{H}{16},\frac{H}{16},512) \rightarrow (\frac{H}{32},\frac{W}{32},512)$ | res-(K-4×4, N-512, P-0, PS, LReLU) |
| | $(\frac{H}{32},\frac{H}{32},512) \rightarrow (\frac{H}{64},\frac{W}{64},512)$ | res-(K-4×4, N-512, P-0, PS, LReLU) |
| | $(\frac{H}{64},\frac{H}{64},512) \rightarrow (\frac{H}{128},\frac{W}{128},512)$ | res-(K-4×4, N-512, P-0, PS, LReLU) |
| | $(\frac{H}{128},\frac{W}{128},512) \rightarrow (\frac{H}{256},\frac{W}{256},512)$ | conv-(K-4×4, N-512, S-2, P-0, PS) |
| U-Net (decoder) | $(\frac{H}{256},\frac{H}{256},512) \rightarrow (\frac{H}{128},\frac{W}{128},512)$ | U + res-(K-4×4, N-512, P-0, PS, IN, ReLU) |
| | $(\frac{H}{128},\frac{H}{128},1024) \rightarrow (\frac{H}{64},\frac{W}{64},512)$ | U + res-(K-4×4, N-512, P-0, PS, IN, ReLU) |
| | $(\frac{H}{64},\frac{H}{64},1024) \rightarrow (\frac{H}{32},\frac{W}{32},512)$ | U + res-(K-4×4, N-512, P-0, PS, IN, ReLU) |
| | $(\frac{H}{32},\frac{H}{32},1024) \rightarrow (\frac{H}{16},\frac{W}{16},512)$ | U + res-(K-4×4, N-512, P-0, PS, IN, ReLU) |
| | $(\frac{H}{16},\frac{H}{16},1024) \rightarrow (\frac{H}{8},\frac{W}{8},256)$ | U + res-(K-4×4, N-256, P-0, PS, IN, ReLU) |
| | $(\frac{H}{8},\frac{H}{8},512) \rightarrow (\frac{H}{4},\frac{W}{4},128)$ | U + res-(K-4×4, N-128, P-0, PS, IN, ReLU) |
| | $(\frac{H}{4},\frac{W}{4},256) \rightarrow (\frac{H}{2},\frac{W}{2},128)$ | U + conv-(K-4×4, N-128, S-1, P-0, PS, IN, ReLU) |
| Decoding | $(\frac{H}{2},\frac{W}{2},128) \rightarrow (H,W,3)$ | residual map: U + conv-(K-4×4, N-3, S-1, P-0, PS, tanh) |
| | $(\frac{H}{2},\frac{W}{2},128) \rightarrow (H,W,1)$ | context mask: U + conv-(K-4×4, N-1, S-1, P-0, PS, sigmoid) |

Table 5.9: Network architecture of dual conditional image generator. We describe each layer or residual block as "conv-(K-, N-, S-, P-, PS/PV, IN/BN, LReLU)", "res(K-, N-, S-, P-, PS/PV, IN/BN, LReLU)". K: kernel size, N: number of filters, S: stride size, P: padding size, PS: padding='same', PV: padding='valid', IN: instance normalisation, BN: batch normalisation, LReLU: LeakyReLU. U: upsampling with kernel size 2×2. Input image size "$H \times W$" is 224×112. Note that the U-Net contains skip connections that are helpful to preserve the underlying image structure across network layers. Downsampling and upsampling residual blocks are depicted in Figure 5.9.

| Part Name | Input → Output Shape | Layer Description |
|---|---|---|
| Input Layer | $(H,W,3) \rightarrow (H,W,3)$ | additive Gaussian noise $\mathcal{N}(0,0.1)$ |
| Hidden Layers | $(H,W,3) \rightarrow (\frac{H}{2},\frac{W}{2},128)$ | conv-(K-4×4, N-128, S-2, P-2, PV, LReLU) |
| | $(\frac{H}{2},\frac{W}{2},128) \rightarrow (\frac{H}{4},\frac{W}{4},256)$ | conv-(K-4×4, N-256, S-2, P-2, PV, IN, LReLU) |
| | $(\frac{H}{4},\frac{W}{4},256) \rightarrow (\frac{H}{4},\frac{W}{4},512)$ | conv-(K-4×4, N-512, S-1, P-2, PV, IN, LReLU) |
| | $(\frac{H}{4},\frac{W}{4},512) \rightarrow (\frac{H}{4},\frac{W}{4},512)$ | conv-(K-4×4, N-512, S-1, P-2, PV, IN, LReLU) |
| Output Layer | $(\frac{H}{4},\frac{W}{4},512) \rightarrow (\frac{H}{4},\frac{W}{4},1)$ | conv-(K-4×4, N-1, S-1, P-2, PV, sigmoid) |

Table 5.10: Network architecture of domain discriminator $D_d$.

| Part Name | Input → Output Shape | Layer Description |
|---|---|---|
| Hidden Layers | $(H,W,3) \rightarrow (\frac{H}{2},\frac{W}{2},64)$ | conv-(K-4×4, N-64, S-2, P-1, PV, LReLU) |
| | $(\frac{H}{2},\frac{W}{2},64) \rightarrow (\frac{H}{4},\frac{W}{4},128)$ | conv-(K-4×4, N-128, S-2, P-1, PV, BN, LReLU) |
| | $(\frac{H}{4},\frac{W}{4},128) \rightarrow (\frac{H}{8},\frac{W}{8},256)$ | conv-(K-4×4, N-256, S-2, P-1, PV, BN, LReLU) |
| | $(\frac{H}{8},\frac{W}{8},256) \rightarrow (\frac{H}{16},\frac{W}{16},512)$ | conv-(K-4×4, N-512, S-2, P-1, PV, BN, LReLU) |
| Pooling Layer | $(\frac{H}{32},\frac{W}{32},512) \rightarrow (1,1,512)$ | average-pooling & dropout=0.999 |
| Output Layer | $(1,1,512) \rightarrow$ C-way softmax | conv-(K-1×1, N-C, S-2, softmax) |

Table 5.11: Network architecture of camera discriminator $D_{cam}$.

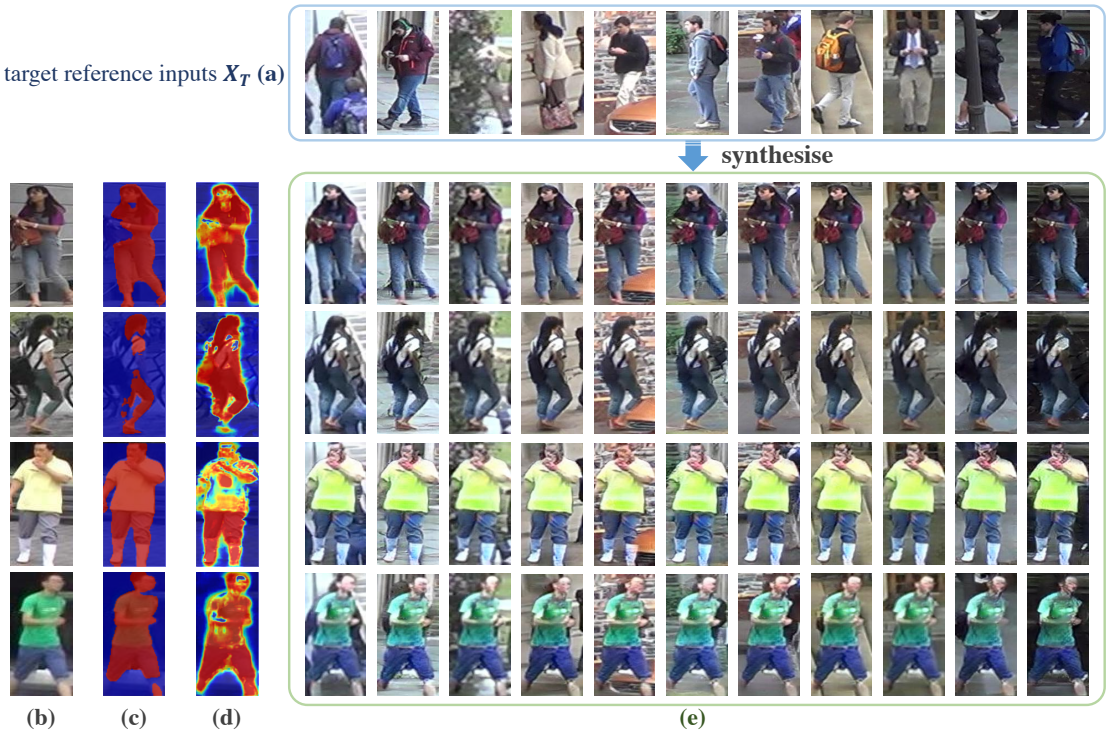



Figure 5.10: Synthetic data by CR-GAN on Market1501→DukeMTMCreID. $X_S$: source image; $X_T$: target image; $M_S$: parsing mask of $X_S$; $1 - X_C$: the inverse of context mask; $X_G$: generated image.


Figure 5.11: Synthetic data by CR-GAN on CUHK03 → DukeMTMCreID. $X_S$: source image; $X_T$: target image; $M_S$: parsing mask of $X_S$; $1 - X_C$: the inverse of context mask; $X_G$: generated image.

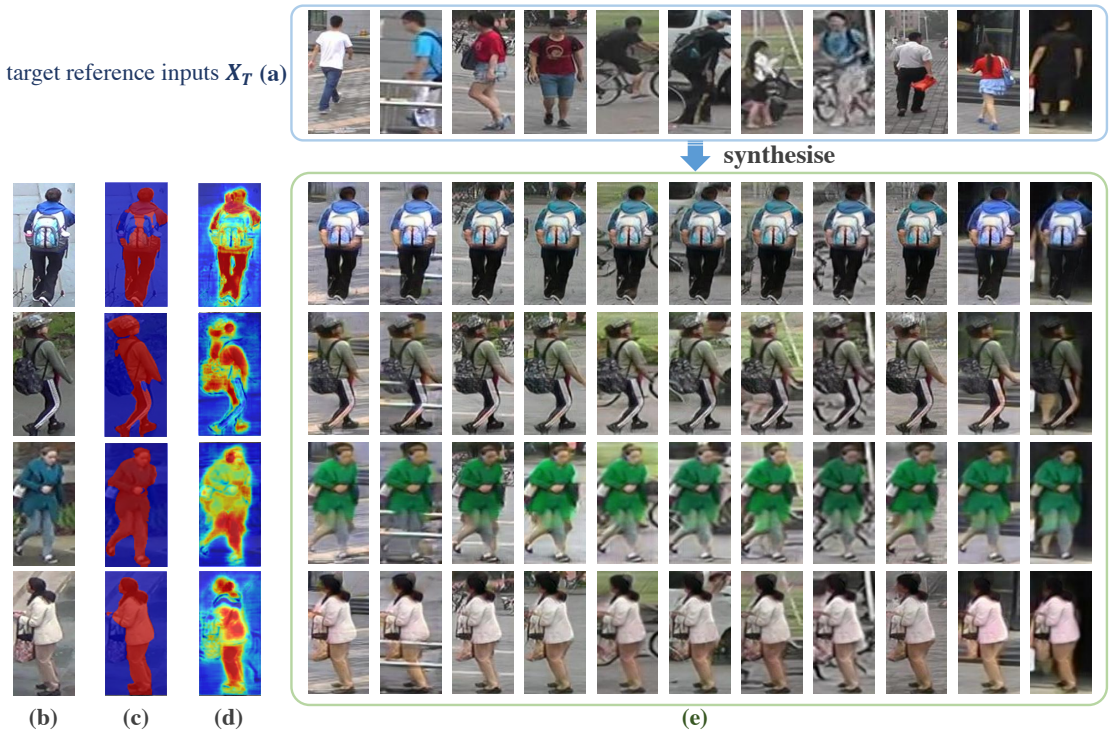



Figure 5.12: Synthetic data by CR-GAN on DukeMTMCreID → Market1501. $X_S$: source image; $X_T$: target image; $M_S$: parsing mask of $X_S$; $1 - X_C$: the inverse of context mask; $X_G$: generated image.

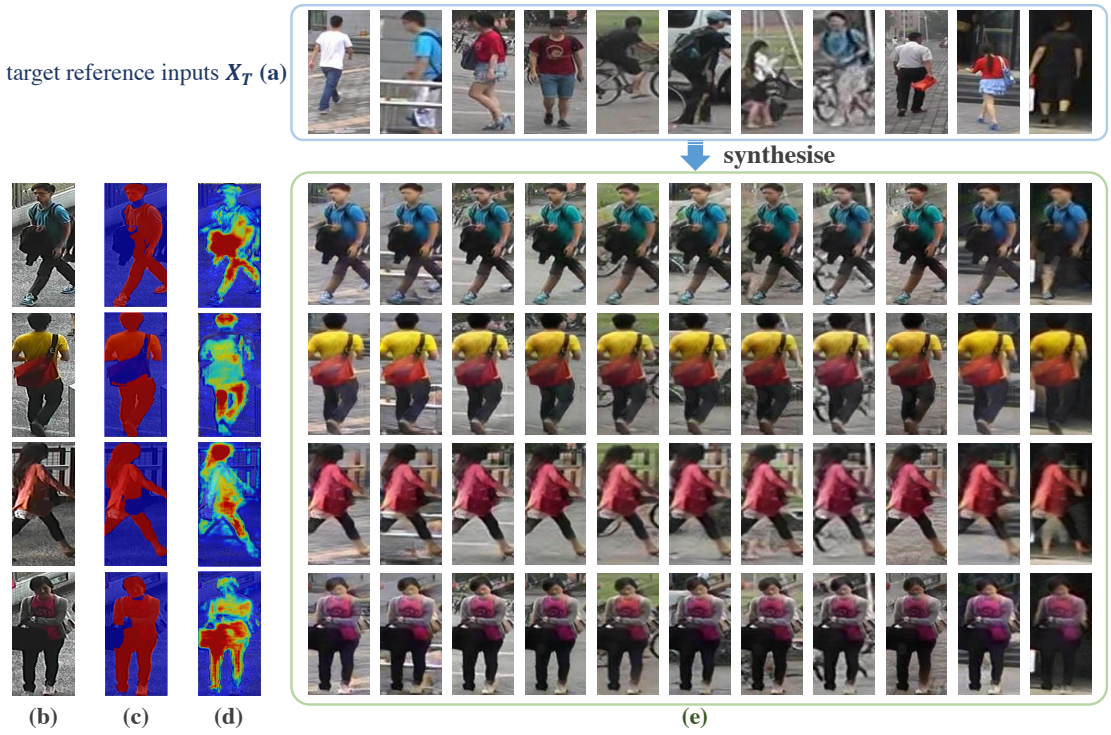



Figure 5.13: Synthetic data by CR-GAN on CUHK03 → Market1501. $X_S$: source image; $X_T$: target image; $M_S$: parsing mask of $X_S$; $1 - X_C$: the inverse of context mask; $X_G$: generated image.

# Chapter 6

# Unsupervised Learning by Online Deep Association

---

Person re-identification (re-id) aims to match persons across disjoint camera views distributed at different locations [Gong et al., 2014], which is a practical computer vision task for real-world surveillance systems. In such real-world surveillance systems, manual annotation is not only prohibitively expensive to collect in practice as there are a quadratic number of camera pairs, but also implausible in many cases due to no sufficient training people reappearing under every pair of camera views. Unsupervised video-based person re-id, therefore, is a non-trivial task as it targets to address person re-id using unlabelled video data without any manual labelling efforts.

While most recent re-id methods rely on static images [Li et al., d; Ahmed et al.; Xiao et al., 2016; Wang et al., 2016a; Li et al., 2017; Sun et al., a; Zheng et al., c; Chen et al., 2017c; Li et al., 2018; Zhong et al., 2018; Wang et al., 2018a; Zhu et al., 2017], video-based re-id has gained increasing attention [Hirzer et al., 2011; Wang et al., a, 2016b; Zhu et al., 2016; Zheng et al., a; You et al., 2016; McLaughlin et al., 2016; Yan et al.; Zheng et al., a; Zhou et al., 2017; Xu et al., 2017] due to the rich space-time information inherently carried in the video tracklets. A video tracklet is a sequence of images that captures rich variations of the same person in terms of occlusion, background clutter, viewpoint, human poses, etc, which can naturally be used as informative data sources for unsupervised learning in person re-id. The majority of current techniques in video person re-id consider the supervised learning context, which imposes a strong assumption on the availability of identity (ID) labels for every camera pair, therefore allowing more powerful and discriminative re-id models to be learned when given relatively small-sized training data. However, supervised learning methods are weak in scaling to real-

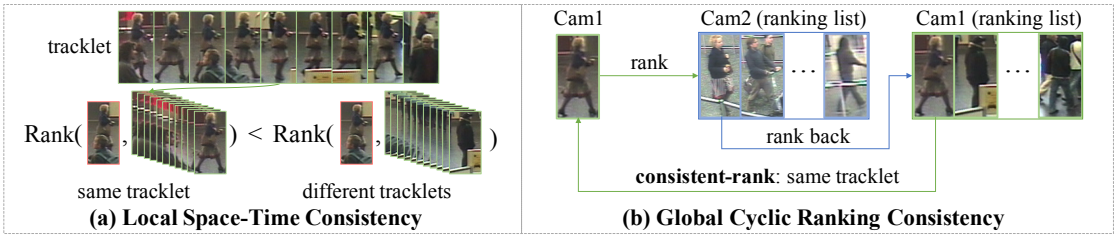



Figure 6.1: Two types of consistency in our Deep Association Learning scheme. (a) Local space-time consistency: Most images from the same tracklet generally depict the same person. (b) Global cyclic ranking consistency: Two tracklets from different cameras are highly associated if they are *mutually* the nearest neighbour returned by a cross-view ranking.

world deployment beyond the labelled training data domains. In practice, exhaustive manual annotation at every camera pair is not only prohibitively expensive for a large identity population across a large camera network, but it is also implausible due to insufficient designated persons reappearing in every camera pair in real-world surveillance context. In this regard, unsupervised video re-id is a more realistic task that is worth studying to improve the scalability of re-id models in practical use.

**Unsupervised learning** methods [Ma et al., 2017; Liu et al., b; Ye et al.; Liu et al., a; Karanam et al.; Wang et al., 2018b] are particularly essential when the re-id task needs to be performed on a large amount of unlabelled video surveillance data cumulated continuously over time, whilst the pairwise ID labels cannot be easily acquired for supervised model learning. Due to the inherent nature of unsupervised learning, existing methods suffer from significant performance degradations when compared to supervised learning methods in video person re-id. For instance, the state-of-the-art rank-1 re-id matching rate on MARS [Zheng et al., a] is only 36.8% by unsupervised learning [Ye et al.], as compared to 82.3% by supervised learning [Li et al., c]. In fact, even the latest video-based unsupervised learning models [Liu et al., b; Ye et al.] for person re-id still lack a principled mechanism to explore the more powerful representation-learning capabilities of deep Convolutional Neural Networks (CNNs) [Bengio et al., 2013] for jointly learning an expressive embedding representation and a discriminative re-id matching model in an end-to-end manner. It is indeed not straightforward to formulate a deep learning scheme for unsupervised video-based person re-id due to: (1) The general supervised learning nature of deep CNN networks: most deep learning objectives are formulated on labelled training data; (2) The cross-camera variations of the same-ID tracklet pairs from disjoint camera views and the likelihood of different people being visually similar in public space, which collectively render the nearest-neighbour distance measure unreliable to capture the cross-view person identity match-

ing for guiding the model learning.

In this work, we aim to tackle the task of unsupervised video person re-id by an end-to-end optimised deep learning scheme without utilising any identity labels. Towards this aim, we formulate a novel *unsupervised* **Deep Association Learning** (DAL) scheme designed specifically to explore two types of *consistency*, including (1) *local space-time consistency* within each tracklet from the same camera view, and (2) *global cyclic ranking consistency* between tracklets across disjoint camera views (Figure 6.1). In particular, we define two margin-based association losses, with one derived from the intra-camera tracklet representation updated incrementally on account of the *local space-time consistency*, and the other derived from the cross-camera representation learned continuously based on the *global cyclic ranking consistency*. Importantly, this scheme enables the deep model to start with learning from the local consistency, whilst incrementally self-discovering more cross-camera highly associated tracklets subject to the global consistency for progressively enhancing discriminative feature learning. Our key idea is to associate the frame-level feature representations to the the nearest neighbours of tracklet-level feature representations within and across camera views. This is based on the *consistency* assumption [Zhou et al., 2004] that the nearest neighbours are likely to belong to the same class. Overall, our DAL scheme imposes batch-wise self-supervised learning cycles to eliminate the need for manual labelled supervision in the course of model training.

In summary, our **contribution** is three-fold:

- We propose for the first time an end-to-end deep learning scheme for unsupervised video person re-id without imposing any human knowledge on identity information.

- We formulate a novel *Deep Association Learning* (DAL) scheme, with two discriminative association losses derived from (1) *local space-time consistency* within each tracklet and (2) *global cyclic ranking consistency* between tracklets across disjoint camera views. Our DAL loss formulation allows typical deep CNNs to be readily trained by standard stochastic gradient descent algorithms.

- Extensive experiments demonstrate the advantages of DAL over the state-of-the-art unsupervised video person re-id methods on three benchmark datasets: PRID2011 [Hirzer et al., 2011], iLIDS-VID [Wang et al., a], and MARS [Zheng et al., a].

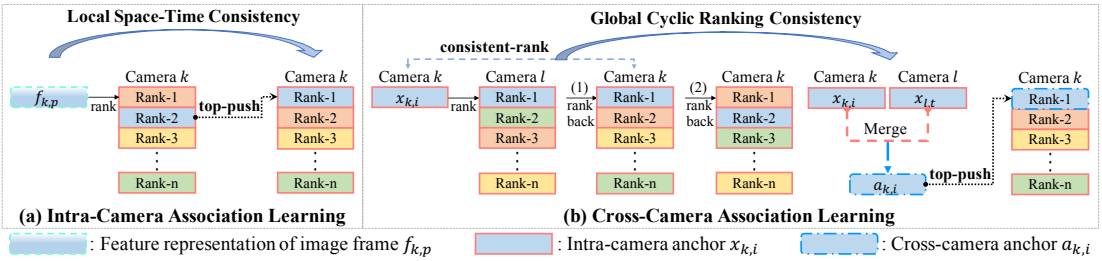



Figure 6.2: Illustration of Deep Association Learning: (a) Intra-camera association learning based on the local space-time consistency within tracklets (Section 6.2). (2) Cross-camera association learning based on the global cyclic ranking consistency on cross-camera tracklets (Section 6.3). Best viewed in colour.

## 6.1 Deep Association Learning

Our goal is to learn a re-id matching model to discriminate the appearance difference and reliably associate the video tracklets across disjoint camera views without utilising any ID labels. Towards this goal, we propose a novel *Deep Association Learning* (DAL) scheme that optimises a deep CNN model based on the learning objective derived based on two types of consistency. As illustrated in Figure 6.2, we explore the *local space-time consistency* and *global cyclic ranking consistency* to formulate two top-push margin-based association losses. In particular, two sets of "anchors" are gradually learned all along the training process for our loss formulation. They are **(1)** a set of *intra-camera anchors* $\{x_{k,i}\}_{i=1}^{N_k}$ that denote the intra-camera feature representations of $N_k$ tracklets under camera $k$; and **(2)** a set of *cross-camera anchors* $\{a_{k,i}\}_{i=1}^{N_k}$, with each representing the cross-camera feature representation merged by the intra-camera feature representations of two highly associated tracklets from disjoint camera views. Overall, the DAL scheme consists of two batch-wise iterative procedures: **(a)** intra-camera association learning and **(b)** cross-camera association learning, as elaborated in the following.

## 6.2 Intra-Camera Association Learning

Intra-camera association learning aims at discriminating intra-camera video tracklets. To this end, we formulate a top-push margin-based intra-camera association loss in the form of the hinge loss based on the ranking relationship of each image frame in association to all the video tracklets from the same camera view. This loss is formulated in three steps as follows.

### 6.2.1 Learning Intra-Camera Anchors

On account of the *local space-time consistency* as depicted Figure 6.1, each video tracklet can simply be represented as a univocal sequence-level feature representation by utilising certain temporal pooling strategy, such as max-pooling or mean-pooling [McLaughlin et al., 2016; Zheng et al., a]. This, however, is time-consuming to compute at each mini-batch learning iteration, as it requires to feed-forward all image frames of each video tracklet through the deep model. To overcome this problem, we propose to represent a tracklet from camera $k$ as an *intra-camera anchor* $x_{k,i}$, which is the intra-camera tracklet representation incrementally updated by the frame representation $f_{k,p}$ of any constituent image frame from the same source tracklet all through the training process. Specifically, the exponential moving average (EMA) strategy is adopted to update each anchor $x_{k,i}$ as follows.

$$x_{k,i}^{t+1} \leftarrow x_{k,i}^{t} - \eta\left(\ell_2(x_{k,i}^{t}) - \ell_2(f_{k,p}^{t})\right), \text{ if } i = p \tag{6.1}$$

where $\eta$ refers to the update rate (set to 0.5), $\ell_2(\cdot)$ is $\ell_2$ normalisation (i.e. $\|\ell_2(\cdot)\|_2 = 1$), and $t$ is the mini-batch learning iteration. As $x_{k,i}$ is initialised as the mean of the frame representations for each tracklet and incrementally updated as Eq. (6.1), the intra-camera anchor is consistently learned all along with the model learning progress to represent each tracklet.

### 6.2.2 Tracklet Association Ranking

Given the set of incrementally updated *intra-camera anchors* $\{x_{k,i}\}_{i=1}^{N_k}$ for camera $k$, the ranking relationship of the frame representation $f_{k,p}$ in association to all intra-camera anchors from the same camera $k$ can be generated based on pairwise similarity measure. We use the $\ell_2$ distance to measure the pairwise similarities between an in-batch frame representation $f_{k,p}$ and all the intra-camera anchor $\{x_{k,i}\}_{i=1}^{N_k}$. Accordingly, a ranking list is obtained by sorting the pairwise similarities of $f_{k,p}$ w.r.t. $\{x_{k,i}\}_{i=1}^{N_k}$, with the rank-1 (top-1) intra-camera anchor having the minimal pairwise distance:

$$\{D_{p,i}|D_{p,i} = \left\|\ell_2(f_{k,p}) - \ell_2(x_{k,i})\right\|_2, i \in N_k\} \xrightarrow{\text{ranking}} D_{p,t} = \min_{i\in[1,N_k]} D_{p,i} \tag{6.2}$$

where $\{D_{p,i}\}_{i=1}^{N_k}$ is the set of pairwise distances between $f_{k,p}$ and $\{x_{k,i}\}_{i=1}^{N_k}$; while $D_{p,t}$ denotes the pairwise distance between $f_{k,p}$ and the rank-1 tracklet $x_{k,t}$.

### 6.2.3 Intra-Camera Association Loss

Given the ranking list for the frame representation $f_{k,p}$ (Eq. (6.2)), the intra-camera rank-1 tracklet $x_{k,t}$ should ideally correspond to the source tracklet $x_{k,p}$ that contains the same constituent frame due to the *local space-time consistency*. We therefore define a top-push margin-based intra-camera association loss to enforce proper association of each frame to the source tracklet for discriminative model learning:

$$\mathcal{L}_I = \begin{cases} [D_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\ [D_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} \end{cases} \tag{6.3}$$

where $[\cdot]_+ = \max(0, \cdot)$, $D_{p,p}$ is the pairwise distance between $f_{k,p}$ and $x_{k,p}$ (the source tracklet), $\overline{D_{j,t}} = \frac{1}{M}\sum_{j=1}^{M} D_{j,t}$ is the averaged rank-1 pairwise distance of the $M$ sampled image frames from camera $k$ in a mini-batch. $m$ is the margin that enforces the deep model to assign the source tracklet as the top-rank. More specifically, if the rank-1 is not the source tracklet (i.e. $p \neq t$), $\mathcal{L}_I$ will correct the model by imposing a large penalty to push the source tracklet to the top-rank. Otherwise, $\mathcal{L}_I$ will further minimise the intra-tracklet variation w.r.t. the averaged rank-1 pairwise distance in each mini-batch. Since $\mathcal{L}_I$ is computed based on the sampled image frames and the up-to-date intra-camera anchors in each mini-batch, it can be efficiently optimised by the standard stochastic gradient descent to adjust the deep CNN parameters iteratively. Overall, $\mathcal{L}_I$ encourages to learn the discrimination on intra-camera tracklets for facilitating the more challenging cross-camera association, as described next.

## 6.3 Cross-Camera Association Learning

A key of video re-id is to leverage the cross-camera ID pairing information for model learning. However, such information is missing in unsupervised learning. We overcome this problem by self-discovering the cross-camera tracklet association in a progressive way during model training. To permit learning expressive representation invariant to the cross-camera appearance variations inherently carried in associated tracklet pairs from disjoint camera views, we formulate another top-push margin-based intra-camera association loss in the same form as Eq. (6.3). Crucially, we extend the tracklet representation to carry the information of cross-camera appearance variations by incrementally learning a set of *cross-camera anchors*. This intra-camera association loss is formulated in three steps as below.

### 6.3.1 Cyclic Ranking

Given the incrementally updated intra-camera anchors (Eq. (6.1)), we propose to exploit the underlying relations between tracklets for discovering the association between tracklets across different cameras. Specifically, a cyclic ranking process is conducted to attain the pair of highly associated intra-camera anchors across cameras as follows.

$$\begin{array}{c} x_{k,i} \xrightarrow{\text{ranking in cam } l} Dc_{p,t} = min_{i\in[1,N_l]} \xrightarrow{\text{ranking back in cam } k} Dc_{q,j} = min_{i\in[1,N_k]} \\ \text{Global Cyclic Ranking Consistency: } j{=}i \end{array} \tag{6.4}$$

where $Dc_{p,t}$ denotes the cross-camera pairwise distance between two intra-camera anchors: $x_{k,p}$ from camera $k$ and $x_{l,t}$ from another camera $l$. Both $Dc_{p,t}$ and $Dc_{q,j}$ denote the rank-1 pairwise distance. The pairwise distance and the ranking are computed same as Eq. (6.2). With Eq. (6.4), we aim to discover the most associated intra-camera anchors across cameras under the criterion of *global cyclic ranking consistency*: $x_{k,p}$ and $x_{l,t}$ are mutually the rank-1 match pair to each other when one is given as a query to search for the best-matched intra-camera anchor in the other camera view. This cyclic ranking process is conceptually related to the cycle-consistency constraints formulated to enforce the pairwise correspondence between similar instances [Zhou et al., 2016; Sener et al.; Godard et al., 2017]. In particular, our *global cyclic ranking consistency* in this process aims to exploit the mutual consistency induced by transitivity for discovering the highly associated tracklets across disjoint camera views all along the model training process.

### 6.3.2 Learning Cross-Camera Anchors

Based on *global cyclic ranking consistency*, we define the cross-camera representation as a *cross-camera anchor* $a_{k,i}$ by merging two highly associated intra-camera anchors as depicted in Figure 6.2 and detailed below.

$$a_{k,i}^{t+1} \leftarrow \begin{cases} \frac{1}{2}\left(\ell_2(x_{k,i}^{t+1}) + \ell_2(x_{l,t}^{t})\right), & \text{if } j = i \quad \text{(Cyclic ranking consistent)} \\ x_{k,i}^{t+1}, & \text{others} \end{cases} \tag{6.5}$$

where $a_{k,i}$ is simply a counterpart of $x_{k,i}$. Each cross-camera anchor is updated as the arithmetic mean of two intra-camera anchors if the consistency condition is fulfilled (i.e. $j = i$), otherwise as the same intra-camera anchor. As the deep model is updated continuously to discriminate the appearance difference among tracklets, more intra-camera anchors are progressively discovered to be highly associated. That is, all along the training process, more cross-camera anchors are

gradually updated by merging the highly associated intra-camera anchors to carry the information of cross-camera appearance variations induced by the tracklet pairs that come from disjoint camera views but potentially depict the same identities.

### 6.3.3 Cross-Camera Association Loss

Given the continuously updated *cross-camera anchors* $\{a_{k,i}\}_{i=1}^{N_k}$, we define another top-push margin-based cross-camera association loss in the same form as Eq. (6.3) to enable learning from cross-camera appearance variations:

$$\mathcal{L}_C = \begin{cases} [Da_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\ [Da_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} \end{cases} \tag{6.6}$$

where $Da_{p,p}$ denotes the pairwise distance between the frame representation $f_{k,p}$ and the cross-camera anchor $a_{k,p}$. Both $D_{p,t}$ and $\overline{D_{j,t}}$ are the same quantities as $\mathcal{L}_I$ in Eq. (6.3). As depicted in Figure 6.2 and in the same spirit as $\mathcal{L}_I$, the cross-camera association loss $\mathcal{L}_C$ enforces the deep model to push the best-associated cross-camera anchor as the top-rank, so as to align the frame representation $f_{k,p}$ towards the corresponding cross-camera representation.

## 6.4 Model Training

**Overall Learning Objective.** The final learning objective for DAL is to jointly optimise two association losses (Eq. (6.3), (6.6)) as follows.

$$\mathcal{L}_{DAL} = \mathcal{L}_I + \lambda \mathcal{L}_C \tag{6.7}$$

where $\lambda$ is a tradeoff parameter that is set to 1 to ensure both loss terms contribute equally to the learning process. The margin $m$ in both Eq. (6.3) and Eq. (6.6) is empirically set to 0.2 in our experiments. The algorithmic overview of model training is summarised in Algorithm 4.

**Complexity Analysis.** We analyse the per-batch per-sample complexity cost induced by DAL. In association ranking (Eq. (6.2)), the pairwise distances are computed between each in-batch image frame and $N_k$ intra-camera anchors for each camera, which leads to a computation complexity of $\mathcal{O}(N_k)$ for distance computation and $\mathcal{O}(N_k log(N_k))$ for ranking. Similarly, in cyclic ranking (Eq. (6.4)), the total computation complexity is $\mathcal{O}(N_l + N_k) + \mathcal{O}(N_l log(N_l) + N_k log(N_k))$. All the distance measures are simply computed by matrix manipulation on GPU with single floating point precision for computational efficiency.

**Algorithm 4** Deep Association Learning.

**Input:** Unlabelled video tracklets captured from different cameras.

**Output:** A deep CNN model for re-id matching.

**for** $t = 1$ **to** *max_iter* **do**

Randomly sample a mini-batch of image frames.

Network forward propagation.

Tracklet association ranking on the *intra-camera anchors* (Eq. (6.2)).

Compute two margin-based association loss terms (Eq. (6.3), (6.6)).

Update the corresponding *intra-camera anchors* based on the EMA strategy (Eq. (6.1)).

Update the corresponding *cross-camera anchors* based on cyclic ranking (Eq. (6.4), (6.5)).

Network update by back-propagation (Eq. (6.7)).

**end for**

**Remark.** Our proposed approach is overall formulated based on the *consistency* assumption, which considers that the two nearest neighbours are likely to belong to the same class. As for intra-camera learning, each frame representation is considered to belong to its tacklet representation. As for inter-camera learning, a cyclic ranking consistency is imposed to discover mutual nearest neighbours across two camera views; while the nearest neighbours are considered to come from the same class. It is possible that the two nearest neighbours may belong to different classes. However, given the pairs of similar nearest neighbours are more likely to belong to the same class than the pairs of dissimilar non-nearest neighbours, our margin-based hinge loss would not associate the very dissimilar pairs from different classes, thus ensuring to learn meaningful discriminative representations.

## 6.5 Discussion

The commonality of most existing unsupervised learning methods in video person re-id is to discover the matching correlations between tracklets across cameras. For example, Ma et al. [Ma et al., 2017] formulate a time shift dynamic warping model to automatically pair cross-camera tracklets by matching partial segments of each tracklet generated over all time shifts. Ye et al. [Ye et al.] propose a dynamic graph matching method to mine the cross-camera labels for iteratively learning a discriminative distance metric model. Liu et al. [Liu et al., b] develop a stepwise metric learning method to progressively estimate the cross-camera labels; but it requires

stringent video filtering to obtain one tracklet per ID per camera for discriminative model initialisation. Our proposed Deep Association Learning (DAL) method in this work differs significantly from previous works in three aspects: **(I)** Unlike [Ma et al., 2017; Liu et al., b], our DAL does not require additional manual effort to select tracklets for model initialisation, which results in better scalability to large-scale video data. **(II)** All existing methods rely on a good external feature extractor for metric learning; while our DAL jointly learns a re-id matching model with discriminative representation in a fully end-to-end manner. **(II)** Our DAL uniquely utilises the intra-camera local space-time consistency and cross-camera global cyclic ranking consistency to formulate the learning objective with a relatively low computational cost.

In essence, our proposed DAL method performs unsupervised **deep metric learning** by learning a nonlinear mapping that transforms input images into a feature representation space, in which the distances within the same class are enforced to be small whilst the distances between different classes are maintained large. Although a variety of supervised deep distance metric learning methods have been proposed to solve the task of person re-id [Li et al., d; Yi et al., 2014; Ahmed et al.; Ding et al., 2015; Liu et al., 2016; Wang et al., 2016a; Cheng et al., 2016; Chen et al., 2016; McLaughlin et al., 2016; Chen et al., 2017b; Hermans et al., 2017; Xu et al., 2017], such as using Siamese network with a pairwise similarity measure objective [Yi et al., 2014; McLaughlin et al., 2016; Xu et al., 2017], or minimising a margin-based hinge loss with a batch online mining strategy for triplet generation [Ding et al., 2015; Hermans et al., 2017], our DAL especially learns a deep embedding representation in an unsupervised fashion. In particular, instead of grounding the learning objective based on pairwise or triple-wise comparison between a few labelled samples, e.g., two samples as a pair or three samples as a triplet, our DAL uniquely learns two set of anchors as the intra-camera and cross-camera tracklet representations, which allows to measure the pairwise similarities between each image frame and all the other tracklet representations to ensure more comprehensive comparison between samples, thus resulting in more effective unsupervised learning.

## 6.6 Experiments

### 6.6.1 Evaluation on Unsupervised Video Person Re-ID

**Datasets.** We conduct extensive experiments on three video person re-id benchmark datasets, including PRID 2011 [Hirzer et al., 2011], iLIDS-VID [Wang et al., a] and MARS [Zheng et al.,

(a) PRID 2011 (b) iLIDS-VID (c) MARS

Figure 6.3: Example pairs of tracklets from three benchmark datasets. Cross-camera variations include changes in illumination, viewpoints, resolution, occlusion, background clutter, human poses, etc.

a] (Figure 6.3). The PRID 2011 dataset contains 1,134 tracklets captured from two disjoint surveillance cameras with 385 and 749 tracklets from the first and second cameras. Among all video tracklets, 200 persons are captured in both cameras. The iLIDS-VID dataset includes 600 video tracklets of 300 persons. Each person has 2 tracklets from two non-overlapping camera views in an airport arrival hall. The MARS has a total of 20,478 tracklets of 1,261 persons captured from a camera network with 6 near-synchronized cameras at a university campus. All the tracklets were automatically generated by the DPM detector [Felzenszwalb et al., 2009] and the GMMCP tracker [Dehghan et al.].

**Evaluation Protocols.** For PRID 2011, following [Wang et al., a; Ye et al.; Liu et al., b] we use the tracklet pairs from 178 persons, with each tracklet containing over 27 frames. These 178 persons are further randomly divided into two halves (89/89) for training and testing. For iLIDS-VID, all 300 persons are also divided into two halves (150/150) for training and testing. For both datasets, we repeat 10 random training/testing ID splits as [Wang et al., a] to ensure statistically stable results. The average Cumulated Matching Characteristics (CMC) are adopted as the performance metrics. For MARS, we follow the standard training/testing split [Zheng et al., a]: all tracklets of 625 persons for training and the remaining tracklets of 636 persons for testing. Both the averaged CMC and the mean Average Precision (mAP) are used to measure re-id performance on MARS. Note, our method does not utilise any ID labels for model initialisation or training.

**Implementation Details.** We implement our DAL scheme in Tensorflow [Abadi et al., 2016]. To evaluate its generalisation ability of incorporating with different network architectures, we adopt two standard CNNs as the backbone networks: ResNet50 [He et al., 2016] and MobileNet [Howard et al., 2017]. Both deep models are initialised with weights pre-trained on ImageNet [Deng et al., 2009]. On the small-scale datasets (PRID 2011 and iLIDS-VID), we apply the RMSProp optimiser [Tieleman and Hinton, 2012] to train the DAL for $2\times10^4$ iterations,

| Datasets | PRID 2011 | | | | iLIDS-VID | | | | MARS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank@$k$ | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | mAP |
| DVDL | 40.6 | 69.7 | 77.8 | 85.6 | 25.9 | 48.2 | 57.3 | 68.9 | - | - | - | - | - |
| STFV3D | 42.1 | 71.9 | 84.4 | 91.6 | 37.0 | 64.3 | 77.0 | 86.9 | - | - | - | - | - |
| MDTS-DTW | 41.7 | 67.1 | 79.4 | 90.1 | 31.5 | 62.1 | 72.8 | 82.4 | - | - | - | - | - |
| UnKISS | 59.2 | 81.7 | 90.6 | 96.1 | 38.2 | 65.7 | 75.9 | 84.1 | - | - | - | - | - |
| DGM+IDE | 56.4 | 81.3 | 88.0 | 96.4 | 36.2 | 62.8 | 73.6 | 82.7 | 36.8 | 54.0 | 61.6 | 68.5 | 21.3 |
| Stepwise | 80.9 | 95.6 | 98.8 | 99.4 | 41.7 | 66.3 | 74.1 | 80.7 | 23.6 | 35.8 | - | 44.9 | 10.5 |
| **DAL (ResNet50)** | 85.3 | 97.0 | 98.8 | 99.6 | 56.9 | 80.6 | 87.3 | 91.9 | 46.8 | 63.9 | 71.6 | 77.5 | 21.4 |
| **DAL (MobileNet)** | 84.6 | 96.3 | 98.4 | 99.1 | 52.8 | 76.7 | 83.4 | 91.6 | 49.3 | 65.9 | 72.2 | 77.9 | 23.0 |

Table 6.1: Evaluation on three benchmarks in comparison to the state-of-the-art unsupervised video re-id methods. Red: the best performance. Blue: the second best performance. '-': no reported results.

with an initial learning rate of 0.045 and decayed exponentially by 0.94 every 2 epochs. On the large-scale dataset (MARS), we adopt the standard stochastic gradient descent (SGD) to train the DAL for $1\times10^5$ iterations, with an initial learning rate of 0.01 and decayed to 0.001 in the last $5\times10^4$ iterations. The batch size is all set to 64. At test time, we obtain the tracklet representation by max-pooling on the image frame features followed by $\ell_2$ normalisation. We compute the $\ell_2$-distance between the cross-camera tracklet representations as the similarity measure for the final video re-id matching.

**Comparison to state-of-the-art methods.** We compare DAL against six state-of-the-art video-based unsupervised re-id methods: DVDL [Karanam et al.], STFV3D [Liu et al., a], MDTS-DTW [Ma et al., 2017], UnKISS [Khan and Bremond, 2016], DGM+IDE [Ye et al.], and Stepwise [Liu et al., b]. Among all methods, DAL is the only unsupervised deep re-id model that is optimised in an end-to-end manner. Table 6.1 shows a clear performance superiority of DAL over all other competitors on the three benchmark datasets. In particular, the rank-1 matching accuracy is improved by 4.4%(85.3-80.9) on PRID 2011, 15.2%(56.9-41.7) on iLIDS-VID and 12.5%(49.3-36.8) on MARS. This consistently shows the advantage of DAL over existing methods for unsupervised video re-id due to the joint effect of optimising two association losses to enable learning feature representation invariant to cross-camera appearance variations whilst discriminative to appearance difference. Note, the strongest existing model DGM+IDE [Ye et al.] additionally uses ID label information from one camera view for model initialisation, whilst Stepwise [Liu et al., b] assumes one tracklet per ID per camera by implicitly using ID labels.

| Datasets | PRID 2011 | | | | iLIDS-VID | | | | MARS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank@$k$ | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | mAP |
| $\mathcal{L}_I$ Only | 62.7 | 85.7 | 92.1 | 96.7 | 31.7 | 55.2 | 67.5 | 78.6 | 41.6 | 59.0 | 66.2 | 73.2 | 16.8 |
| $\mathcal{L}_C$ Only | 81.6 | 95.2 | 98.1 | 99.7 | 47.4 | 72.6 | 81.5 | 89.2 | 48.1 | 65.3 | 71.4 | 77.6 | 22.6 |
| $\mathcal{L}_I$+$\mathcal{L}_C$ | 84.6 | 96.3 | 98.4 | 99.1 | 52.8 | 76.7 | 83.4 | 91.6 | 49.3 | 65.9 | 72.2 | 77.9 | 23.0 |

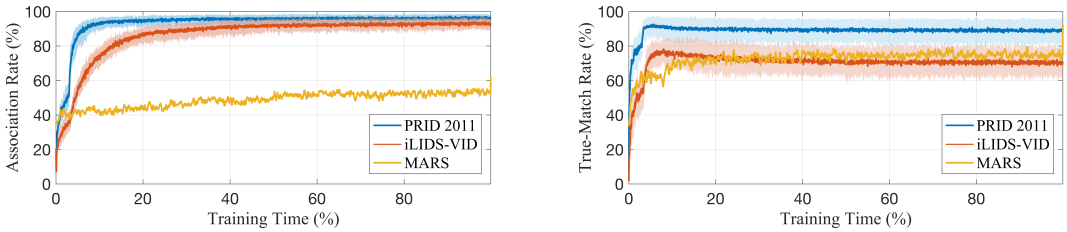
Table 6.2: Effectiveness of two association losses. Red: the best performance. CNN: MobileNet.

In contrast, DAL uses neither of such additional label information for model initialisation or training. More crucially, DAL consistently produces similar strong re-id performance with different network architectures (ResNet50 and MobileNet), which demonstrates its applicability to existing standard CNNs.

### 6.6.2 Component Analyses and Further Discussion

**Effectiveness of two association losses.** The DAL trains the deep CNN model based on the joint effect of two association losses: (1) intra-camera association loss $\mathcal{L}_I$ (Eq. (6.3)) and (2) cross-camera association loss $\mathcal{L}_C$ (Eq. (6.3)). We evaluate the individual effect of each loss term by eliminating the other term from the overall learning objective (Eq. (6.7)). As shown in Table 6.2, jointly optimising two losses leads to the best model performance. This indicates the complementary benefits of the two loss terms in discriminative feature learning. Moreover, applying $\mathcal{L}_C$ alone has already achieved better performance as compared to the state-of-the-art methods in Table 6.1. When comparing with $\mathcal{L}_I$+$\mathcal{L}_C$, applying $\mathcal{L}_C$ alone only drop the rank-1 accuracy by 3.0%(84.6-81.6), 5.4%(52.8-47.4), 1.2%(49.3-48.1) on PRID 2011, iLIDS-VID, MARS respectively. This shows that even optimising the cross-camera association loss *alone* can still yield competitive re-id performance, which owes to its additional effect in enhancing cross-camera invariant representation learning by reliably associating tracklets across disjoint camera views all along the training process.

**Evolution of cross-camera tracklet association.** As aforementioned, learning representation robust to cross-camera variations is a key to learning an effective video re-id model. To understand the effect of utilising the cyclic ranking consistency to discover highly associated tracklets during training, we track the proportion of *cross-camera anchors* that are updated to denote the cross-camera representation by merging two highly associated tracklets (*intra-camera anchors*). Figure 6.4(a) shows that on PRID 2011 and iLIDS-VID, 90+% tracklets find their highly associ-




(a) Evolution on association rate.

(b) Evolution on true-match rate.

Figure 6.4: Evolution on cross-camera tracklet association. The shaded areas denote the varying range of 10-split results repeated on PRID 2011 and iLIDS-VID. Best viewed in colour.

| Datasets | PRID 2011 | | | | iLIDS-VID | | | | MARS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank@$k$ | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 | mAP |
| DAL ($\mathcal{L}_I$+$\mathcal{L}_C$) | 84.6 | 96.3 | 98.4 | 99.1 | 52.8 | 76.7 | 83.4 | 91.6 | 49.3 | 65.9 | 72.2 | 77.9 | 23.0 |
| ID-Supervised | 84.3 | 98.1 | 99.2 | 99.8 | 51.5 | 76.0 | 83.8 | 89.9 | 71.8 | 86.8 | 90.7 | 93.3 | 51.5 |

Table 6.3: Comparison between our unsupervised model and its supervised counterpart. Red: the best performance. CNN: MobileNet.

ated tracklets under another camera at the end of training. On the much noisier large-scale MARS dataset, the DAL can still associate more than half of tracklets (>50%) across cameras. Importantly, as seen in Figure 6.4(b), among self-discovered associated cross-camera tracklet pairs, the percentage of true-match pairs at the end of training is approximately 90% on PRID 2011, 75% on iLIDS-VID, and 77% on MARS, respectively. This shows compellingly the strong capability of DAL in self-discovering the unknown cross-camera tracklet associations without learning from manually labelled data.

**Comparison with supervised counterparts.** We further compare DAL against the supervised counterpart trained using ID labelled data with the identical CNN architecture (MobileNet), denoted as ID-Supervised. This ID-Supervised is trained by the cross-entropy loss computed on the ID labels. Results in Table 6.3 show that: (1) On PRID 2011 and iLIDS-VID, DAL performs similarly well as the ID-Supervised. This is highly consistent with our observations of high tracklet association rate in in Figure 6.4, indicating that discovering more cross-camera highly associated tracklets can help to learn a more discriminative re-id model that is robust to cross-camera variations. (2) On MARS, there is a clear performance gap between the supervised and unsupervised models. This is largely due to a relatively low tracklet association rate arising from the difficulty of discovering cross-camera tracklet associations in a larger identity population among much noisier tracklets, as indicated in Figure 6.4(a).

## 6.7 Summary

In this work, we presented a novel *Deep Association Learning* (DAL) scheme for unsupervised video person re-id using unlabelled video tracklets extracted from surveillance video data. Our DAL permits deep re-id models to be trained without any ID labelling for training data, which is therefore more scalable to deployment on large-sized surveillance video data than supervised learning based models. In contrast to existing unsupervised video re-id methods that either require more stringent one-camera ID labelling or per-camera tracklet filtering, DAL is capable of learning to automatically discover the more reliable cross-camera tracklet associations for addressing the video re-id task without utilising ID labels. This is achieved by jointly optimising two margin-based association losses formulated based on the *local space-time consistency* and *global cyclic ranking consistency*. Extensive comparative experiments on three video person re-id benchmarks show compellingly the clear advantages of the proposed DAL scheme over a wide variety of state-of-the-art unsupervised video person re-id methods. We also provided detailed component analyses to further discuss the insights on how each part of our method design contributes towards the overall model performance.

# Chapter 7

# Conclusion and Future Work

---

Deep learning algorithms are known to be data hungry. To achieve human- or even super-human-level performance in most visual recognition tasks, large collections of labelled datasets are generally required to formulate meaningful supervision signals for model optimisation. The lack of label annotations is therefore a major lingering issue that prevents us to build more generalisable visual recognition models. To mitigate the need of tremendous label annotations, this thesis studied four different visual learning paradigms in limited-label regime, all of which share the same aim to address the challenge of lacking sufficient label annotations in visual recognition tasks, including visual classification and visual search. Specifically, our key idea is to exploit abundant unlabelled visual data together with limited labelled data, by optimising an auxiliary unsupervised learning signal formulated without utilising any label annotations. In the sequel, a conclusion of this thesis and a discussion on future work are further elaborated.

## 7.1 Conclusion

In conclusion, this thesis has explored and tackled different visual learning paradigms in limited-label regime, where each learning paradigm considers the unlabelled data samples are presented under different conditions (see Figure 1 in Abstract). As outlined in Chapter 1 and further analysed in Chapter 2, according to **(i)** whether there exists class mismatch between the labelled set and unlabelled set, **(ii)** whether there exists domain drift between the labelled set and unlabelled set, and **(iii)** whether the labelled set is available, **visual learning in limited-label regime** can be further divided into four different learning paradigms, which are presented in

Chapter 3, 4, 5, 6 and summarised below.

1. Chapter 3 addressed close-set semi-supervised learning (i.e. no class mismatch or domain drift) by introducing a memory module augmented to the network during training. The memory module captures the underlying manifold structure to reliably propagate probabilistic class assignments from the labelled data pool to the unlabelled data samples, thus allowing to learn from the unlabelled data without any label information.

2. Chapter 4 tackled an under-explored visual learning paradigm: open-set semi-supervised learning (i.e. with class mismatch) by formulating an uncertainty-aware self-distillation (UASD) scheme. UASD averages all the historical model predictions during training to derive less overconfident probabilistic class assignments (a.k.a. soft target) on the unlabelled data, which could be leveraged to discard samples lying out-of-distribution and selectively propagate reliable label assignments on unlabelled samples to avoid catastrophic error propagation.

3. Chapter 5 studied a practical and challenging visual learning paradigm: open-set cross-domain learning (i.e. class mismatch and domain drift) by forming a novel instance-guided context rendering scheme. A dual conditional image generation framework is built to render the source person identity population into the target domain contexts, thus producing an abundant amount of synthetic imagery data for domain-adaptive training. By fine-tuning upon these synthetic data, the learnt representations become discriminative to person identity labels and invariant to the contextual variations, thus yielding a more robust visual search model in person re-identification.

4. Chapter 6 investigated unsupervised learning (i.e. none labelled data) in visual search via a new deep association learning (DAL) scheme. DAL learns from unlabelled videos by automatically associating image representations to video representations via a within-view temporal consistency loss and a cross-view cycle consistency loss, both of which collectively result in more discriminative representations for effective visual search in person re-identification.

From Chapter 2 to Chapter 6, it is demonstrated that visual learning in limited-label regime can be investigated and solved under different scenarios, which introduce the unlabelled visual data in different ways. Although these visual learning paradigms are mainly evaluated on visual

classification and visual search tasks, the underlying generic modelling principles to *propagate*, *selectively propagate*, *transfer* and *discover* label assignments on the unlabelled visual data could be easily generalised to other visual recognition tasks.

It is worth noting that our proposed methods are proposed to tackle different problem scenarios in a case-by-case basic. In fact, due to unique data characteristics of the unlabelled data in different scenarios, it is impractical to design one universal semi-supervised or unsupervised learning model for solving all kinds of visual learning problems in the limited-label regime. For instance, in the close-set semi-supervised learning scenario, where we know that the unlabelled data is drawn from the same class distribution as the labelled data, all the unlabelled data can be used for semi-supervised training. In the open-set semi-supervised learning scenario, where we have the labelled data and unlabelled data sampled from very different label spaces, it is essential to use the only in-distribution unlabelled data for model training while discarding the out-of-distribution unlabelled data. In the open-set cross-domain learning scenario with labelled data and unlabelled data sampled from very different label spaces and domains, it is important to tackle both the label drift and domain drift to ensure an effective utilisation of the unlabelled data. In the most extreme case without any labelled data (i.e. unsupervised learning), it is essential to discover the label information automatically. While the former two scenarios are more commonly seen in object classification task (with a known label space), the latter two scenarios are more common in visual recognition tasks (where the label annotations in an open label space could not be easily acquired). In conclusion, the different data characteristics in various problem scenarios suggest the necessity to address the challenges on a case-by-case basis. Notably, our proposed methods are formulated to address these unique problem-specific challenges.

In the following, future directions are further discussed to shed light on potential extension for our studied learning paradigms.

## 7.2 Future Work

Visual learning meaningful representations with limited or none human supervision is a long-standing problem. It is therefore profound to formulate effective auxiliary supervision signals that empower machine learning models to learn from visual data without relying on task-relevant manual annotations provided in the form of instance-level class labels (e.g. visual classification), pixel-level class labels (e.g. semantic segmentation), bounding boxes (e.g. object detection),

etc. In essence, such auxiliary supervision signals aim to introduce additional *regularisers* that constrain the model solution space in a meaningful manner, which in effect learn invariant representations that are robust to various task-irrelevant invariances while being discriminant to task-relevant factors. In addition to semi-supervised and unsupervised visual learning paradigms for improving model generalisation, there are also other plausible learning paradigms to serve with a similar purpose. In the following, we discuss the possible extensions and highlight the scope for further development beyond the works presented in this thesis.

**Learning from Transferable Knowledge** entails the possibility to accrue better model generalisation without utilising numerous label annotations for model training. The generic aim is to leverage prior knowledge learnt from another data distribution by introducing auxiliary unsupervised supervision signals based on prior information encoded in *pre-trained model parameters* [Sharif Razavian et al.; Yosinski et al.; Oquab et al.; Wang et al., b], *model representations* [Donahue et al., 2014; Romero et al., 2015], or *model predictions* [Hinton et al., 2015; Laine and Aila, 2017; Ba and Caruana; Urban et al.]. For instance, to transfer model parameters, a new deep model architecture can be built by adding new layers upon an existing model [Wang et al., b]. To transfer model representations or predictions, a new model can be optimised with an auxiliary regulariser formed upon representations or predictions from a pre-trained larger model or an ensemble of models [Romero et al., 2015; Hinton et al., 2015]. As learning from transferable knowledge allows us to utilise label information from other datasets, it can also be easily extended to tackle the lack of numerous label annotations. The plausible extension in visual learning is to leverage the pre-trained model parameters, representations, or predictions to impose auxiliary unsupervised loss terms that permit to learn from unlabelled visual data. During the course of this thesis, we also proposed a *feature regularisation* technique by *consensus propagation* [Chen et al., 2017c], which transfers the knowledge learnt from a multi-scale teacher model to a single-scale student model. Our proposed deep multi-scale representation learning model is a natural extension upon standard teacher-student transfer learning [Hinton et al., 2015], which can be easily extended to transfer knowledge from labelled data to unlabelled data in a semi-supervised or unsupervised manner.

**Learning Using Privileged Information** (LUPI) [Vapnik and Vashist, 2009; Vapnik and Izmailov, 2015] is originally proposed by Vapnik et al. as a machine learning paradigm to use *extra side information available only during training*, which could also be exploited for visual

learning in limited-label regime. This learning paradigm has further been introduced to a variety of computer vision tasks (e.g. image retrieval [Sharmanska et al., 2013] and web image recognition [Li et al., e]) to enhance model performance by exploiting additional information sources [Sharmanska et al., 2013; Li et al., e; Lopez-Paz et al., 2015; Hoffman et al., 2016; Yang et al., 2017; Garcia et al.; Lambert et al., 2018; Lee et al., b], such as supervision provided by other modalities [Hoffman et al., 2016; Garcia et al.], and auxiliary hashtags [Sharmanska et al., 2013; Li et al., e]. In the case of lacking sufficient label supervision, LUPI provides an alternative way to formulate auxiliary supervision signals that could further boost model generalisation without labelling more visual data to meet the expected model performance. For instance, Hoffman et al. [Hoffman et al., 2016] use depth images as extra data source to guide the learning of more effective RGB image representation for object detection. Lee et al. [Lee et al., b] propose a GAN framework that uses labelled stimulated synthetic images as privilege information to constrain the learning on unlabelled real-world images for semantic segmentation in urban scenes. During the course of this thesis, we also proposed an auxiliary visual-semantic optimisation scheme to tie the learnt visual representation with auxiliary language semantics extracted from privileged linguistic information, such as image captions or product descriptions that carry semantic information about the visual content [Chen et al., 2020a]. As a natural future exploration, our proposed visual-semantic optimisation scheme can be extended for visual learning in limited-label regime by exploiting privileged information to form auxiliary supervision signals that do not urge for task-specific manual annotations.

**Learning from Non-Visual Data** provides a promising modelling strategy to learn from non-vision modalities, such as audio, language, and touch, which can be formulated as a multi-modal semi-supervised or unsupervised learning paradigm to learn visual representations without labelling the visual data. In the domain of vision-audio learning, based upon the natural synchronisation between vision and audio in a video, unsupervised learning signals can be simply formed to enforce the visual-audio consistency across two modalities [Korbar et al.; Owens and Efros]. By applying unsupervised clustering within the vision and audio modalities, single-modal and multi-modal pseudo labels can be generated to guide representation learning in both modalities [Alwassel et al., 2019; Morgado et al., 2020]. In the domain of vision-language learning, built upon recent advance of BERT in unsupervised linguistic representation learning [Devlin et al., 2018], variants of visual-linguistic BERT have been proposed to learn the joint representations

of vision and language in an unsupervised manner [Tan and Bansal, 2019; Lu et al.; Su et al.; Li et al., a], which could serve as effective unsupervised pre-training strategies without using any label annotations to improve a variety of downstream tasks, such as image captioning, image-text matching, and visual question answering. In essence, multi-modal data could serve as informative sources to provide richer unsupervised signals for semi-supervised or unsupervised learning. A natural future extension beyond visual learning is to explore multi-modal learning in limited-label regime by leveraging additional non-visual modalities to provide informative unsupervised learning signals and guide representation learning with minimal human supervision.

# Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

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 Representation*, Vancouver, B.C., Canada, April 2018.

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representation*, New Orleans, Louisiana, USA, May 2019.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*.

Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Mahsa Baktashmotlagh, Masoud Faraki, Tom Drummond, and Mathieu Salzmann. Learning factorized representations for open-set domain adaptation. In *International Conference on Learning Representation*.

Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. In *International Conference on Learning Representation*.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

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

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory*, 1998.

Avrim Blum, John Lafferty, Mugizi Robert Rwebangira, and Rajashekar Reddy. Semi-supervised learning using randomized mincuts. In *International Conference on Machine Learning*, Alberta, Canada, July 2004.

TE Boult, S Cruz, AR Dhamija, M Gunther, J Henrydoss, and WJ Scheirer. Learning and the unknown: Surveying steps toward open world recognition. In *aaai Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, February 2019.

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*, Honolulu, Hawaii, USA, July 2017.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representation*.

Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *IEEE International Conference on Computer Vision*.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, a.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *IEEE International Conference on Computer Vision*, b.

O Chapelle, A Zien, Cowell Z Ghahramani, et al. Semi-supervised classification by low density separation. In *International Workshop on Artificial Intelligence and Statistics*, 2005.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *European Conference on Computer Vision*.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017a.

Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

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, July 2017b.

Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval.

Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al. Person re-identification by deep learning multi-scale representations. In *Workshop of IEEE International Conference on Computer Vision*, Venice, Italy, October 2017c.

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

Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *European Conference on Computer Vision*, Munich, Germany, September 2018b.

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, October 2019a.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seatle, USA, June 2020a.

Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *aaai Conference on Artificial Intelligence*, New York City, USA, February 2020b.

Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019b.

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.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pages 561–580. Springer, 2012.

Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in Neural Information Processing Systems*, Long Beach, California, USA, December 2017.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, Florida, USA, June 2009.

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*, Salt Lake City, Utah, USA, June 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *The Conference on Association for Computational Linguistics*, Melbourne, Australia, July 2018.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision*.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representation*.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, Atlanta, USA, June 2014.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision*.

Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018.

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*, 32(9):1627–1645, 2009.

Rob Fergus, Yair Weiss, and Antonio Torralba. Semi-supervised learning in gigantic image collections. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2009.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, New York City, USA, June 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, Lille, France, July 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision*.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. In *British Machine Vision Conference*, London, UK, September 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representation*.

Herbert P Ginsburg and Sylvia Opper. *Piaget's theory of intellectual development*. Prentice-Hall, Inc, 1988.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing.

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

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*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2005.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, May 2010.

Philip Haeusser, Alexander Mordvintsev, and Daniel Cremers. Learning by association-a versatile semi-supervised training method for neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seatle, USA, June 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representation*, New Orleans, Louisiana, USA, May 2019.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, Ystad Saltsjöbad, Sweden, May 2011.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, Stockholm, Sweden, July 2018.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE International Conference on Computer Vision*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *The Conference on Uncertainty in Artificial Intelligence*, Monterey, California, USA, August 2018.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, Bled, Slovenia, June 1999.

Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V

Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *IEEE International Conference on Computer Vision*.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Furqan M Khan and Francois Bremond. Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Colorado Springs, CO, USA, August 2016.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2014.

Piotr Koniusz, Yusuf Tas, and Fatih Porikli. Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Harrahs and Harveys, Lake Tahoe, USA, December 2012.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, Long Beach, California, USA, December 2017.

John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. at&t labs, 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop of International Conference on Machine Learning*, Atlanta, USA, June 2013.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, a.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representation*, Vancouver, B.C., Canada, April 2018a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018b.

Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. In *International Conference on Learning Representation*, b.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *aaai Conference on Artificial Intelligence*, a.

Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision*, b.

Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, c.

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*, d.

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.

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.

Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *European Conference on Computer Vision*, e.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representation*, Vancouver, B.C., Canada, April 2018.

Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Junyong Zhu. M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. In *aaai Conference on Artificial Intelligence*.

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*.

Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

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. In *British Machine Vision Conference*, a.

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*, b.

Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, June 2019.

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 *ACM International Conference on Multimedia*, Amsterdam, Netherlands, October 2016.

Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *IEEE International Conference on Computer Vision*, a.

Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *IEEE International Conference on Computer Vision*, b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, Lille, France, July 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representation*, May 2015.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*.

Andy J Ma, Jiawei Li, Pong C Yuen, and Ping Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE transactions on image processing*, 24(5):1599–1613, 2015.

Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In *International Conference on Machine Learning*, New York City, USA, June 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2019.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*.

Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, November 2016.

Tom M Mitchell. Generalization as search. *Artificial intelligence*, 18(2):203–226, 1982.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. In *International Conference on Learning Representation*, San Juan, Puerto Rico, May 2016.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:2004.12943*, 2020.

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015.

Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Conference on Information and Knowledge Management*, 2000.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level

image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representation*, Toulon, France, April 2017.

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.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representation*, Toulon, France, April 2017.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Marc'Aurelio Ranzato and Martin Szummer. Semi-supervised learning of compact document representations with deep networks. In *International Conference on Machine Learning*, Helsinki, Finland, July 2008.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2015.

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*, Montréal, Canada, December 2015.

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*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representation*, May 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, October 2015.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *IEEE Workshops on Application of Computer Vision*, pages 29–36, 2005.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. Technical report, 1985.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, Crete, Greece, September 2010.

Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision*.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

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.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, New York City, USA, June 2016.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015.

Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*.

Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*.

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013.

Mingguang Shi and Bing Zhang. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics*, 27(21):3017–3023, 2011.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *International Conference on Learning Representation*, San Juan, Puerto Rico, May 2016.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representation*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2015.

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

Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision*, a.

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*, b.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representation*, Banff, Canada, April 2014.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representation*.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *The Conference on Association for Computational Linguistics*, Florence, Italy, July 2019.

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.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, pages 26–31, 2012.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, Araucano Park, Las Condes, Chile, December 2015.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, June 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Gregor Urban, Krzysztof J. Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Rich Caruana, Abdel rahman Mohamed, Matthai Philipose, and Matthew Richardson. Do deep convolutional nets really need to be deep (or even convolutional)? In *International Conference on Learning Representation*.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research*, 16(1):2023–2049, 2015.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, Helsinki, Finland, July 2008.

Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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 2016a.

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, 2018a.

Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018b.

Qin Wang, Wen Li, and Luc Van Gool. Semi-supervised learning by augmented distribution alignment. In *IEEE International Conference on Computer Vision*, Seoul, Korea, October 2019.

Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, a.

Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016b.

Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(8): 1447–1460, 2015.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*.

Yuxiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *IEEE Conference on Computer Vision and Pattern Recognition*, b.

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.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, Amsterdam, Netherlands, October 2016.

Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *International Conference on Machine Learning*, Helsinki, Finland, July 2008.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representation*, Banff, Canada, April 2014.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seatle, USA, June 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

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.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, Stockholm, Sweden, July 2018.

Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*.

Hao Yang, Joey Tianyi Zhou, Jianfei Cai, and Yew Soon Ong. Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *IEEE International Conference on Computer Vision*.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *IEEE International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014.

Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision*, Amsterdam, Netherlands, October 2016.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*.

Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, York, UK, September 2016.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *IEEE International Conference on Computer Vision*, Seoul, Korea, October 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018a.

Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018b.

Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at alibaba. In *International Conference on Knowledge Discovery & Data Mining*, pages 993–1001, London, United Kingdom, August 2018c.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *European Conference on Computer Vision*.

Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, a.

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*, b.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, c.

Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *European Conference on Computer Vision*.

Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2004.

Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016.

Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, July 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*, Washington D.C, USA, August 2003.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, and Hui Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *International Joint Conference of Artificial Intelligence*, New York, United States, July 2016.

Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 27(5):2286–2300, 2017.