Unsupervised Deep Learning of Visual Representations

Jiabo Huang

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

School of Electronic Engineering and Computer Science

Queen Mary University of London

13 October 2022

Unsupervised Deep Learning of Visual Representations

Jiabo Huang

Abstract

Interpreting visual signals from complex imagery and video data with a few or no human annotation is challenging yet essential for realising the true values of deep learning techniques in real-world scenarios. During the past decade, deep learning has achieved unprecedented breakthroughs in lots of computer vision fields. Nonetheless, to optimise a large number of parameters in deep neural networks for deriving complex mappings from an input visual space to a discriminative feature representational space, the success of deep learning is heavily relying on a massive amount of human-annotated training data. Collecting such manual annotations are labour-intensive, especially in large-scale that has been proven to be critical to learning generalisable models applicable to new and unseen data. This dramatically limits the usability and scalability of deep learning when being applied in practice. This thesis aims to reduce the reliance of learning deep neural networks on exhaustive human annotations by proposing novel algorithms to learn the underlying visual semantics with *insufficient/inadequate* manual labels, denoted as *generalised* unsupervised learning. Based on the different assumptions on the available sources of knowledge used for learning, this thesis studies generalised unsupervised deep learning from four perspectives including learning without any labels by knowledge aggregation from local data structure and knowledge discovery from global data structure, transferring knowledge from relevant labels, and propagating knowledge from incomplete labels. Specifically, novel methods are introduced to address unresolved challenges in these problems as follows:

Chapter 3 The first problem is *aggregating knowledge from local data structure*, which assumes that apparent visual similarities (pixel intensity) among images are encoded in local neighbourhoods in a feature representational space, providing partially the underlying semantic relationships among samples. This thesis studies discriminative representation learning in this problem, aiming to derive visual features which are discriminative in terms of image's semantic class memberships. This problem is challenging because it is scarcely possible without ground-truth labels to accurately determine reliable neighbourhoods encoding the same underlying class concepts, considering the arbitrarily complex appearance patterns and variations both within and across classes. Existing methods learning from hypothetical inter-sample relationships tend to be error-propagated as the incorrect pairwise supervisions are prone to accumulate across the training process and impact the learned representations. To that end, this thesis proposes to progressively discover sample anchored / centred neighbourhoods to reason and learn the underlying semantic relationships among samples iteratively and accumulatively. Moreover, a novel progressive affinity diffusion process is presented to propagate reliable inter-sample relationships across adjacent neighbourhoods, so as to further identify the within-class visual variation from between-class similarity and bridge the gap between low-level imagery appearance (*e.g.* pixel intensity) and high-level semantic concepts (*e.g.* object class memberships).

Chapter 4 The second problem is *discovering knowledge from global data structure*, which makes an assumption that visual similarity among samples of the same semantic classes is gen-

erally higher than that of different classes. This thesis investigates deep clustering for solving this problem which simultaneously learns visual features and data grouping without any labels. Existing unsupervised deep learning algorithms fails to benefit from joint representations and partitions learning by either overlooking global class memberships (*e.g.* contrastive representation learning) or basing on unreliable pseudo labels estimated by updating feature representations that are subject to error-propagation during training. To benefit clustering of images from discriminative visual features derived by a representation learning process, a Semantic Contrastive Learning method is proposed in this thesis, which concurrently optimises both instance visual similarities and cluster decision boundaries to reason about the hypotheses of semantic classes by their consensus. What's more, based on the observation that assigning visually similar samples into different clusters will implicitly reduce both the intra-cluster compactness and inter-cluster diversity and lead to lower partition confidence, this thesis presents an online deep clustering method named PartItion Confidence mAximisation. It is established on the idea of learning the most semantically plausible data separation by maximising the "global" partition confidence of clustering solution using a novel differentiable partition uncertainty index.

Chapter 5 The third problem is *transferring knowledge from relevant labels*, which assumes the availability of manual labels in relevant domains and the existence of common knowledge shared across domains. This thesis studies transfer clustering in this problem, which aims at learning the semantic class memberships of the unlabelled target data in a novel (target) domain by knowledge transfer from a labelled source domain. Whilst enormous efforts have been made on data annotation during the past decade, accumulating knowledge from existing labelled data to benefit understanding the persistently emerging unlabelled data is intuitively more efficient than exhaustively annotating new data. However, considering the unpredictable changing nature of imagery data distributions, the accumulated pre-learned knowledge does not transfer well without making strong assumptions about the learned source and the novel target domains, *e.g.* from domain adaptation to zero-shot and few-shot learning. To address this problem and effectively transfer knowledge between domains that are different in both data distributions and label spaces, this thesis proposes a self-SUPervised REMEdy method to align knowledge of domains by learning jointly from the intrinsically available relative (pairwise) imagery information in the unlabelled target domain and the prior-knowledge learned from the labelled source domain, so as to benefit from both transfer and self-supervised learning.

Chapter 6 The last problem is *propagating knowledge from incomplete labels*, with the assumption that incomplete labels (*e.g.* collective or inexact) are usually easier to be collected and available but tend to be less reliable. This thesis investigates video activity localisation in this problem to locate a short moment (video segment) in an untrimmed and unstructured video according to a natural language query. To derive discriminative representations of video segments to accurately match with sentences, a temporal annotation of the precise start/end frame indices of each target moments are usually required. However, such temporal labels are not only harder to be collected than pairing videos with sentences as they require carefully going through videos frame-by-frame, but also subject to labelling uncertainty due to the intrinsic ambiguity in a video activity's boundary. To reduce annotation cost for deriving universal visual-textual correlations, a Cross-sentence Relations Mining method is introduced in this thesis to align video segments and query sentences when only a *paragraph* description of activities (collective label) in a video is available but not per-sentence temporal labels. This is accomplished by exploring cross-sentence relationships in a paragraph as constraints to better interpret and match complex moment-wise temporal and semantic relationships in videos. Moreover, this thesis also studies the problem of propagating knowledge to avoid the negative impacts of inexact labels. To that end, an Elastic Moment Bounding method is proposed, which accommodates flexible and adaptive activity temporal boundaries towards modelling universal video-text correlations with tolerance to underlying temporal uncertainties in pre-fixed human annotations.

Figure 1 depicts an overview of the main studies carried out in this thesis.

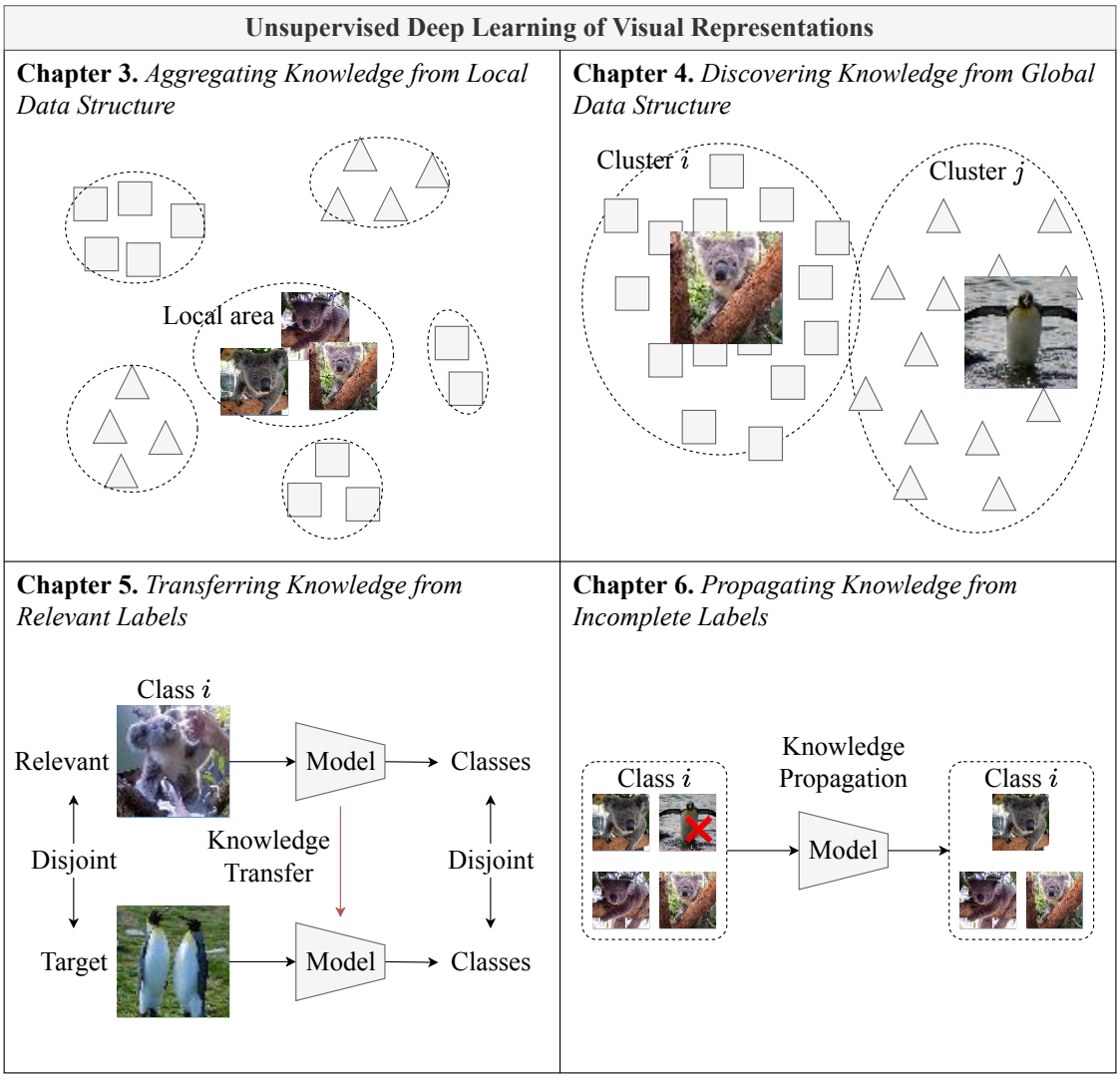


Figure 1: An overview of the main studies carried out in this thesis. According to the different assumptions made as the sources of knowledge for unsupervised deep learning, all studies presented in this thesis are grouped into four categories: aggregating knowledge from local data structure without any labels (**Chapter 3**), discovering knowledge from global data structure without any labels (**Chapter 4**), transferring knowledge from relevant labels (**Chapter 5**), and propagating knowledge from incomplete labels (**Chapter 6**).

Declaration

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

Chapter 3

- **J. Huang**, Q. Dong, S. Gong, X. Zhu. *Unsupervised Deep Learning by Neighbourhood Discovery*. In Proc. International Conference on Machine Learning (ICML), Long beach, CA, USA, June 2019.
- **J. Huang**, Q. Dong, S. Gong, X. Zhu. *Unsupervised Deep Learning via Affinity Diffusion*. In Proc. AAAI Conference on Artificial Intelligence (AAAI), New York, USA, Feb 2020.

Chapter 4

- **J. Huang**, S. Gong, X. Zhu. *Deep Semantic Clustering by Partition Confidence Maximisation*. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, June 2020.
- **J. Huang**, S. Gong. *Deep Clustering by Semantic Contrastive Learning*, Accepted by British Machine Vision Conference (BMVC), London, UK, Nov 2022.

Chapter 5

- **J. Huang**, S. Gong. *Unsupervised Transfer Learning with Self-Supervised Remedy*, arXiv preprint arXiv:2006.04737 (2020).

Chapter 6

- **J. Huang**, Y. Liu, S. Gong, H. Jin. *Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation*. In Proc. IEEE International Conference on Computer Vision (ICCV), Montreal, Canada, Oct 2021.
- **J. Huang**, H. Jin, S. Gong, Y. Liu. *Video Activity Localisation with Uncertainties in Temporal Boundary*, Accepted by European Conference on Computer Vision (ECCV), Tel Aviv, Israel, Oct 2022.

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Prof. Shaogang Gong for his perpetual patience, excellent guidance and enthusiastic supervision. Meanwhile, I convey my special thanks to Dr. Xiatian Zhu for his continued encouragement and invaluable advices. It is through their supervision and guidance that I gradually learn to independently conduct research study and creatively drive research ideas.

I would like to thank Dr. Miles Hansard for being my second supervisor and Dr. Qianni Zhang for being my independent assessor throughout my PhD study. My warm appreciation goes to all members and visiting researchers at QMUL Vision Group for their friendship and support (mostly in chronological order): Dr. Jifei Song, Dr. Qi Dong, Dr. Xiaobing Chang, Dr. Hang Su, Dr. Minxian Li, Dr. Wei Li, Dr. Xu Lan, Dr. Yanbei Chen, Dr. Zhiyi Cheng, Dr. Aytac Kanaci, Dr. Guan'an Wang, Miss Qingze Yin, Dr. Guile Wu, Mr. Pan Li, Mr. Qilei Li, Mr. Weitong Cai, Miss. Shitong Sun, Mr. Jian Hu, Mr. Dezhao Luo, Miss Jiayi Lin and Mr. Ke Han. I am very grateful to my friends from University of Surrey, Dr. Conghui Hu, Dr. Tianyuan Yu, Dr. Anran Qi, Dr. Kaiyue Peng and Dr. Kaiyang Zhou for our friendship. I give my big thanks to my friends Mr. Weihong Li at University of Edinburgh and Mr. Yuedong Chen at Monash University for their consistent and warm encouragement. I am thankful to Dr. Yang Liu at Peking University and Dr. Hailin Jin at Adobe for our fabulous collaboration.

I feel blessed to have the enduring love and endless support from my family members, in particular my parents and my siblings. Meanwhile, I am indebted to Miss Yuwen Shen for her immense love, trust and company all the time.

Last but not least, I give sincere thanks to all friendly QMUL administrative and system support staff for their great help, as well as those who have helped me during my PhD study but I may not have the fortune of knowing their names.

Contents

List of Figures

List of Tables

List of Abbreviations

***k*NN**	*k*-nearest neighbours
ACC	Accuracy
AI	Artificial Intelligence
AN	Anchor Neighbourhood
AND	Anchor Neighbourhood Discovery
ARI	Adjusted Rand Index
ASV	Assignment Statistics Vector
AwA2	Animals with Attributes2 dataset
BCE	Binary Cross Entropy
CD	Cluster Discrimination
CNN	Convolutional Neural Network
CRM	Cross-sentence Relations Mining
CUB	Caltech-UCSD-Birds dataset
DET	Determined boundary
ELA	Elastic boundary
EMB	Elastic Moment Bounding
FC	Fully-connected
FCS	Feature Cosine Similarity
FLO	Oxford Flower dataset
FSL	Few-shot Learning
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HOG	Histogram of Oriented Gradients
HPE	Hard Positive Enhancement

ID	Instance Discrimination
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoU	Intersection over Union
KL	Kullback-Leibler
LC	Linear Classifier
LSTM	Long Short Term Memory
MGIN	Multi-grained Interaction Network
MIL	Multi-Instance Learning
mIoU	Mean Intersection over Union
MMN	Modalities Matching Network
MoI	Moment-of-Interest
NMI	Normalised Mutual Information
OOD	Out-of-Distribution
PAD	Progressive Affinity Diffusion
PICA	PartItion Confidence mAximisation
PUI	Partition Uncertainty Index
SCL	Semantic Contrastive Learning
SCS	Strongly Connected Subgraph
SGD	Stochastic Gradient descent
SMT	Semantic consistency
SOTA	State-of-the-Art
SUN	SUN Attribute dataset
SUPREME	self-SUPervised REMEdy
TMP	Temporal consistency
UDA	Unsupervised Domain Adaptation
ZSL	Zero-shot Learning

Chapter 1

Introduction

1.1 Scope of the Thesis

Visual perception is one of the most essential building bricks of human natural intelligence by perceiving and processing visual information from the surrounding environment to acquire useful knowledge of the world. It is usually referred to as the ability to identify and organise the information captured by human sensors (eye) to convert the resulted neural impulses (visual signals) into a form (visual images) that enables human to interpret the signals and make subsequent decisions, *e.g*. recognising objects, detecting activities, etc. To enlighten machines and endow them the ability to "mimic" human cognition so as to automate tasks, numerous Artificial Intelligence (AI) (Russell and Norvig [155]) researchers have been devoting to the field of computer vision (Sonka et al. [165]) and have developed countless advanced techniques to build artificial visual perception, *i.e*., obtaining high-level understandings of visual information from digital visual data (images and videos) in the forms of decisions. For the sakes of display and storage, visual information is always encoded as a matrix of pixels in computer. However, the independent pixels usually fail to explain a reasonable decision. Therefore, it is critical for artificial visual perception to represent digital visual data more abstractly and holistically by discovering and disentangling the underlying explanatory factors hidden in pixels (Bengio et al. [12]). This thesis studies such a fundamental computer vision task, termed as *visual representation learning* (*a.k.a*. visual feature learning), to map high-dimensional visual data to a compact and discriminative latent feature space, so as to capture the posterior distribution of underlying explanatory

factors for the observations and make it easier to build classifiers or other predictors.

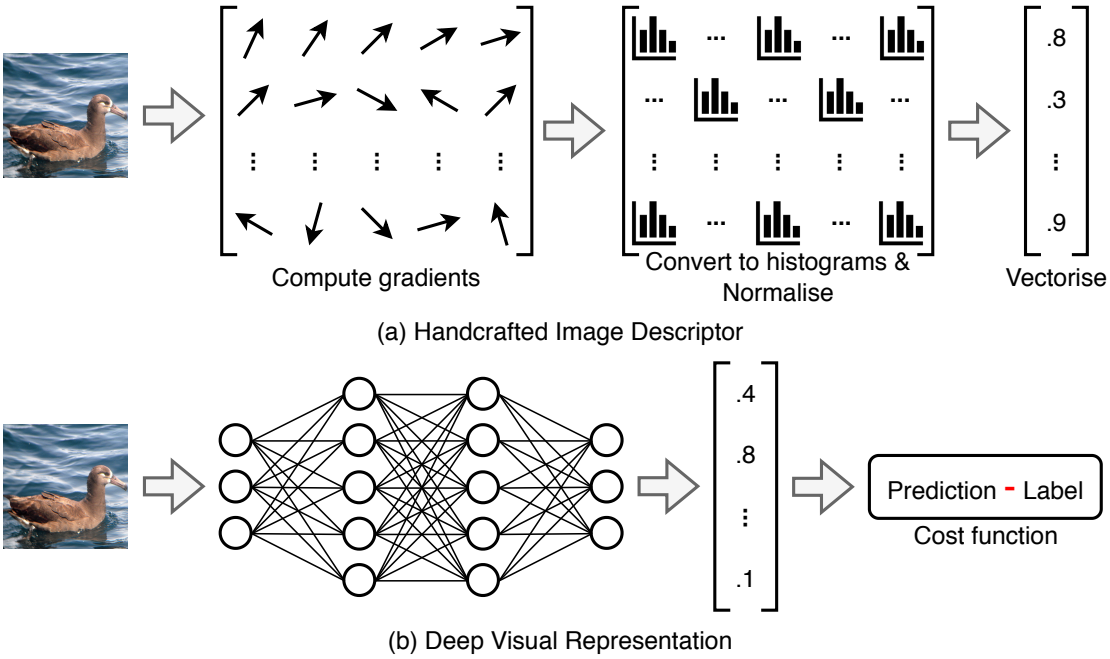


Figure 1.1: An illustration of deep visual representation in comparison to handcrafted image descriptor. We take the Histogram of Oriented Gradients (HOG) (Dalal and Triggs [35]) as an example of handcrafted image descriptors. Whilst **(a)** handcrafted image descriptors were designed by human experts to represent images to be invariant to scale, rotation and other semantic-agnostic visual variations, **(b)** deep visual representations are derived in a data-driven manner by learning a multi-layers neural network according to a task-specific objective loss function to maximise the consistency of network's predictions and human labels.

In the past decade, deep learning (Goodfellow et al. [58]; LeCun et al. [99]) of visual feature representations has made remarkable progress thanks to the enormous efforts put on large-scale datasets collections and annotations as well as the advanced Graphics Processing Units (GPUs). As compared in Figure 1.1, rather than handcrafting substantial descriptors with each aiming at identifying certain types of visual pattern (*e.g.* edge, corner and *etc.*) as the potential explanatory factors for arbitrary tasks according to human expertise (Dalal and Triggs [35]; Lowe et al. [113]; Ojala et al. [133]), deep learning derives feature representations by fitting the mapping functions from the input to the label prediction spaces. This is accomplished by optimising a task-specific objective loss function (*e.g.* cross-entropy for image classification) to maximise the consistency of sample's manual labels and its predictions yielded by a model (neural network (LeCun et al. [99]; Krizhevsky et al. [93])) composed of multiple cascaded neural layers with each conducting a linear or non-linear projection on its own inputs. The term 'deep' in deep learning refers to the large number of layers in a neural network, which has been shown beneficial to modelling

complex mapping functions (He et al. [68]). Given a well-trained neural network model, the outputs of its intermediate layers (usually the second-last layer) can be used to represent the input data tailored for the tasks of interest, encoding essential visual information to make reasonable predictions for unseen samples. This thesis studies *deep learning of visual representations* to discover the underlying explanatory factors of tasks and encode them in the representations of visual data by deep neural networks in a data-driven manner.

Regardless the unprecedented success achieved by deep learning on visual representation learning in recent years, it is heavily relying on the massive amount of visual data with exhaustive and precise manual labels for model training. This restricts its usability and scalability in practice where human annotations are not always available due to the unaffordable labelling cost for large-scale datasets or the lack of task-specific expertise (*e.g*. labels of medical images can only be annotated by human experts (Shen et al. [160])). Without such labels of samples encoding the desired observation-prediction mapping functions, it is clueless for deep learning models to derive expressive and discriminative representations of visual data. In contrast, such exhaustive and precise manual labels are not indispensable for human cognition to understand visual data and describe them conceptually. This is because humans are capable of making use of various sources of knowledge to accomplish new tasks, *e.g*. shared knowledge on related tasks, unreliable knowledge in incomplete labels or even certain intuitions and common senses (assumptions), whilst deep learning models are usually trained and supervised by a single source of knowledge, *i.e*. sample-label pairing relationships. Therefore, to advance artificial visual perception to benefit from persistently emerging unlabelled data and make it applicable in real-world, it is essential to explore the potential of deep learning models to acquire knowledge from different sources other than the task-specific sample-label pairs. To this end, this thesis focuses on *unsupervised deep learning of visual representations* when the human annotations, which implies the desired mapping functions, is insufficient or inadequate during model training.

1.2 Problem Definition

1.2.1 Generalised Unsupervised Learning

By contrast with supervised learning, unsupervised learning (Dayan et al. [36]; Ghahramani [53]; Hinton and Sejnowski [73]; Bengio [9]) seeks for learning the labels of observations with regard to a task of interest without human-annotated outputs associated with each input at the

time of learning. Unsupervised learning is a critical topic to be studied to facilitate artificial visual perception not only because it is commonly exist in human brain (Dayan et al. [36]) but also due to the strong assumption made by supervised learning on the availability of task-specific human annotations that restricts the scalability and deployability of learning effective vision models in real-world. For example, the unaffordable cost of human annotations refrains supervised deep learning from benefiting from larger scale training data. The large scale of data used for model training have been shown to be one of the most essential factors to the success of deep learning (Krizhevsky et al. [93]), especially when the generalisation ability of the deep networks is of concern. However, considering the unaffordable annotation cost, it is impractical to pursue stronger models by scaling up the training datasets with manual labels. Whilst deep neural networks trained on one of the most popular large-scale imagery dataset, ImageNet-1K (Deng et al. [37]) with 1.2 million labelled images from 1,000 natural object classes, have shown notable capacity to generalise to a variety of vision tasks/domains, unsupervised learning have yielded competitive or even superior models by scaling up the training data without needing human annotations or curation in Instagram-1B (Yalniz et al. [205]) composed of $\sim$1 billion public images from Instagram (He et al. [69]). Therefore, to facilitate powerful vision models which are applicable in practice, it is crucial to explore and acquire knowledge beyond human annotations, *e.g.* from the underlying data structure.

This thesis studies unsupervised learning with a focus on discriminative models, which is formally defined as follows:

Definition 1 (Unsupervised Learning) *Given a target problem $\mathcal{P}_t$ composed of visual data from a target domain $\mathcal{X}_t$ and a target task involving a label space $\mathcal{Y}_t$, unsupervised learning aims at deriving a function $f = g(\mathcal{X}_t)$ modelling the underlying mapping from samples to their corresponding labels $\mathcal{Y}_t = f(\mathcal{X}_t)$ with only the observed inputs but not any human annotations.*

In Definition 1, the notion $g(\cdot)$ stands for the deep learning algorithms used to train the deep neural networks $f(\cdot)$ modelling the mapping from visual observations $\mathcal{X}_t$ to predictions ($\mathcal{Y}_t$) of target tasks.

The essence of unsupervised learning is to automate tasks by machine models without depending too much on per-task exhaustive human annotations for training. Conventional unsupervised learning (Definition 1) takes the visual data and its distribution as the only source of knowledge for model learning (Bengio [9]; Dayan et al. [36]) without any kind of human annota-

tions. However, human beings barely learn to solve a new task from scratch by "pure" guestimate but leverage any existing knowledge either accumulated in past or available from the surrounding environment. In other words, there are always a wide variety of knowledge sources beyond the task-specific manual labels and the visual data itself existing in real-world that are ready to benefit visual understanding for novel tasks. In this regard, this thesis investigates and discusses unsupervised learning in a wider context when the machine models can be trained with or without human annotations as long as the annotations are insufficient or inadequate to indicate the underlying mapping functions from the visual inputs to the predictions of target tasks. Such a problem is termed as *generalised unsupervised learning* in this context, which is defined as:

Definition 2 (Generalised Unsupervised Learning) *Given a target problem $\mathcal{P}_t$ composed of visual data from a target domain $\mathcal{X}_t$ and a target task involving a label space $\mathcal{Y}_t$, generalised unsupervised learning $f = g(\mathcal{X}, f')$ seeks for the one-to-one mapping function $\mathcal{Y}_t = f(\mathcal{X}_t)$ from target samples to their labels without the help of the underlying pairwise relationships ($\mathcal{X}_t$, $\mathcal{Y}_t$) but an independent function f' subjecting to $f \neq f'$. The prior mapping function $f' = \mathcal{X}_s \rightarrow \mathcal{Y}_s$ is derived in a source domain $\mathcal{X}_s$ with human annotations for a source task $\mathcal{Y}_s$ being available.*

Comparing to Definition 2, the conventional unsupervised learning given in Definition 1 is a special case when $f' = \emptyset$, *i.e.*, the observed inputs are the only source of knowledge for model learning. By relaxing the constraint on human annotations in Definition 2, the problem of unsupervised learning is considered from a more general perspective and is related to a wider range of techniques that aiming at exploring a variety sources of knowledge beyond the target task-specific manual labels. For example, models in transfer learning[1] (Pan and Yang [134]) derive f' to make use of the knowledge acquired in a source domain and a source task then transfer it across domains (*e.g.* domain adaptation (Patricia and Caputo [135]; Peng et al. [139]; Muandet et al. [122]) where $\mathcal{X}_t \neq \mathcal{X}_s$), across tasks (*e.g.* multi-task learning (Zhang and Yang [222]) where $\mathcal{Y}_t \neq \mathcal{Y}_s$) or across both (*e.g.* few-shot (Sung et al. [168]) and zero-shot (Xian et al. [196]) where $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_s \neq \mathcal{Y}_t$). Moreover, the generalised unsupervised learning is also related to the settings where the source and target problems share both the data distribution ($\mathcal{X}_t = \mathcal{X}_s$) and the label space ($\mathcal{Y}_t = \mathcal{Y}_s$), but the labels are given partially (semi-supervised learning (Chapelle et al. [24])) or inexactly and collectively (weakly-supervised learning (Zhou [230])) to be associated

[1] Whilst there are lots of studies on transfer learning with manual labels being available on the target data and tasks, we discuss it in the context of unsupervised learning in this thesis.

with a set of samples (observations) rather than each individual. Such labels are sometimes easier to be collected but considered inadequate, which may contain a certain amount of sample-label mismatches or biased pairwise relationships. In this case, the function f' encoding the prior-knowledge is still different from the desired mapping function $f \neq f'$, and the pre-learned model f' should be used carefully to refrain f from being misled by unreliable prior.

Problem	Technique	Labels of $\mathcal{X}_s$	Labels of $\mathcal{X}_t$	$\mathcal{X}_s = \mathcal{X}_t$	$\mathcal{Y}_s = \mathcal{Y}_t$
Discovering/Aggregating knowledge from global/local data structure	**Unsupervised Learning**	N/A	None	✓	✓
Transferring knowledge from relevant labels	Domain Adaptation	Exhaustive	None	✗	✓
	Few-shot Learning	Exhaustive	Partial	✗	✗
	Zero-shot Learning	Exhaustive	Collective	✗	✗
	Transfer Learning	Exhaustive	None	✗	✗
Propagating knowledge from incomplete labels	Semi-supervised Learning	N/A	Partial	✓	✓
	Weakly-supervised Learing	N/A	Collective	✓	✓

Table 1.1: Comparisons of different types of generalised unsupervised learning. Given Definition 2, problems related to generalised unsupervised learning can be categorised according to whether and what kinds of **(1)** human annotations (labels of $\mathcal{X}_s$/$\mathcal{X}_t$) are provided **(2)** in the target domains ($\mathcal{X}_s = \mathcal{X}_t$) **(3)** for the target tasks ($\mathcal{Y}_s = \mathcal{Y}_t$). The human annotations in source domain (Labels of $\mathcal{X}_s$) is not applicable when the source and target domains are identical. The problems studied in this thesis are highlighted in **bold**.

1.2.2 Types of Generalised Unsupervised Learning

Given definition 2, the generalised unsupervised learning problem covers a wide range of techniques in deep learning, which aims at comprehensively exploring the available sources of knowledge in real-world to facilitate machine perception. According to where the knowledge used for model training comes from, this thesis studies unsupervised deep learning from four different perspectives, including aggregating knowledge from local data structure, discovering knowledge from global data structure, transferring knowledge from relevant labels, and propagating knowledge from incomplete labels, An elaboration of generalised unsupervised learning types involved in this thesis is given in Table 1.1.

(1) Aggregating knowledge from local data structure seeks to explore the adjacent relationships of unlabelled target data in local neighbourhoods as the knowledge indicating the semantic similarity/consistency of visual observations. The local data structure is defined as

the small group memberships (neighbourhoods in restricted size) of a data collection in this context. Considering the complex patterns and their arbitrary variations and combinations existed in real-world visual data, assuming high visual similarity among all observations with similar semantics may sometimes be over strong and the feature learning process is susceptible to the inevitable errors distributed in clusters. Hence, it is more reasonable to relax such an assumption by holding it only within local areas (neighbourhoods). The knowledge about local adjacent relationships is undeniably less complete than that about global data structure, however, it is usually more reliable (samples of the same groups are more semantically consistent). This enables deep models to explore the holistic cluster memberships with a divide-and-conquer principle either explicitly or implicitly, *i.e*. aggregating knowledge from small groups, so as to obtain semantically discriminative visual representations. In the light of this, this thesis investigates representation learning with the aim of its semantic discrimination capacity by knowledge aggregation.

(2) Discovering knowledge from global data structure aims at mining the knowledge encoded in the cluster memberships that is inherent to a collection of unlabelled target data, where a cluster refers to a potential semantic class of interest, *e.g*. natural object, physical activity, *etc*. The global data structure is considered "inherent" to the data by assuming samples of the same clusters are generally more similar to each other. This problem is closely related to the classical cluster analysis techniques, which aims to separate a set of data in a pre-learned feature space so that each partition is likely to encode certain data characteristics shared by its members, *e.g*. k-means (Coates and Ng [32]), hierarchical clustering (Gowda and Krishna [60]), *etc*. However, the conventional clustering algorithms are usually limited when dealing with high-dimensional visual data as the pixel intensity of images/videos is not discriminative in terms of visual semantics, thus, fails to provide reliable supervision signals to facilitate effective feature learning by deep neural networks. Therefore, it is critical to simultaneously optimise the objectives of representation learning and clustering so to derive discriminative visual features by deep learning models, which is termed as deep clustering (Hershey et al. [72]; Chang et al. [21]) and is the concentration of the second problem studied in this thesis.

(3) Transferring knowledge from relevant labels seeks for the shared knowledge *across different* data domains and/or tasks to benefit the understanding of unlabelled visual ob-

servations in a target domain by the existing knowledge acquired from human annotations in a relevant source domain. This problem is extensively related to a wide range of techniques with different assumptions on the connections of the source and target domains and tasks. As elaborated and compared in Table 1.1, the source and target domains can be linked through a shared tasks (label space) in domain adaptation (Ganin and Lempitsky [49]; Muandet et al. [122]) or aligned by additional human annotations in the target domains, *e.g*. a few labelled target samples (partial labels) in few-shot learning (Ravi and Larochelle [147]) or class-level attribute labels (collective) shared across source and target tasks in zero-shot learning (Ye and Guo [209]). Although those techniques are established with different assumptions to adapt to different scenes, a shared hypothesis held by them is the existence of common knowledge that is agnostic to data domains and tasks. Therefore, in the third problem, this thesis studies knowledge transfer without assuming any relationships of the source and target domains/tasks, which is referred to transfer learning in Table 1.1 and is distinct from other transferring techniques with certain assumption.

(4) Propagating knowledge from incomplete labels aims to obtain visual understanding of the target data with the help of the annotations intendedly collected for the target data itself but missing a reliable indication of the desired mapping function to be modelled. Such a scenario is common in practice, for examples, when a dataset is expanded with new data without annotations (*e.g*. partial labels in semi-supervised learning (Chapelle et al. [24])) or when data is collected from searching engines with certain keywords as their labels (*e.g*. collective and inexact labels in weakly-supervised learning (Zhou [230])). The incomplete labels provide the models with vital cues for understanding the visual observations in the target domain from where the samples of interest (testing samples) are likely drawn, fulfilling the i.i.d. (independent and identically distributed) assumption commonly held in the deep learning community (Tan et al. [172]). However, taking such incomplete labels as oracles for constructing supervision signals for model learning is prone to degrading the generalisation capacity of the resulted models. It is crucial to propagate the knowledge acquired from incomplete labels to the whole dataset in order to yield an unbiased model that is applicable to new and unseen data. This thesis studies the last problem with a focus on knowledge propagation from collective or inexact labels (*i.e*. weakly-supervised learning), where a label is associated with a collection of samples and the knowledge encoded should

be propagated to every individual to derive the universal sample-label mapping functions.

1.3 Challenges and Solutions

This section highlights the unresolved challenges to the four types of generalised unsupervised learning problems, especially those being considered in this thesis (Section 1.2.2), as well as the potential solutions proposed.

1.3.1 Aggregating Knowledge from Local Data Structure

This thesis investigates class discriminative representation learning in the problem of knowledge aggregation from local data structure. In the task of representation learning, unlabelled images are fed into models to produce their feature representations as the only outcomes. The resulted imagery features will then be used to facilitate subsequent object recognition tasks. In recent years, Convolutional Neural Networks (CNNs) trained in a supervised fashion have significantly boosted the state-of-the-art performance of various computer vision tasks (LeCun et al. [98]; Krizhevsky et al. [93]; Simonyan and Zisserman [163]; He et al. [68]). The feature representations yielded by a supervised CNN (*e.g*. trained for classification on ImageNet (Deng et al. [37])) are not only discriminative but also beneficial to novel tasks and/or domains. Despite such remarkable success, this technique is limited due to a number of stringent assumptions. First, supervised model learning requires enormous labelled datasets to be collected manually and exhaustively. This does not always hold valid due to high annotation costs. Besides, the benefits of enlarging labelled datasets may have diminishing returns (Sun et al. [167]). Second, out-of-box deployment [2] of pre-learned deep neural networks across tasks and domains becomes less effective when the target tasks significantly differ from the source tasks (Goodfellow et al. [58]). Due to the only need for accessing unlabelled data which is typically available at scale, unsupervised deep learning provides a conceptually generic and scalable solution to these limitations.

Challenges. The *first* challenge to the problem is to identify the local neighbourhoods containing semantically consistent samples, *i.e*. all the samples falling in the same local area in the latent space are from the same semantic classes. Considering the complex visual patterns existed in real-world visual data as well as their arbitrary combinations, samples from the same semantic classes might be visually distinct to each other due to the changing illumination, occlusions

[2] Out-of-box deployment refers to directly using the trained models on arbitrary new tasks in any new domains without the need to collect data for further fine-tuning.

and *etc*. This will lead to dramatic intra-class visual variations. On the other hand, objects of different semantic concepts may look similar in certain local parts or from some viewpoints, resulting in large inter-class visual similarity. Therefore, the local adjacent relationships between samples don't always reveal their true semantic similarity, especially when the features of samples are yielded by a randomly initiated or poorly trained model. In this case, it is crucial for a model to identify and learn from "pure" local neighbourhoods to get rid of distractions, so as to be sensitive to intra-class discrepancy as well as inter-class affinity. The *second* challenge is finding the trade-off between learning from limited intra-class visual variations in small local neighbourhoods and enlarging neighbourhoods with reduced within-group semantic consistency. Training a model with the knowledge mined from over-large neighbourhoods leads to mostly false-positive inter-sample relationships, while the tiny neighbourhoods tend to yield too many false-negative supervision signals, *e.g*. every single sample is treated as an independent class in instance contrastive learning (Wu et al. [195]; He et al. [69]) for sample specificity learning. In both the cases, the visual features derived is not discriminative regarding their inherent semantics.

Solutions. This thesis presents a generic unsupervised deep learning method for discriminative representation learning called *Anchor Neighbourhood Discovery* (AND). It combines the advantages of both clustering and instance contrastive learning whilst mitigating their disadvantages in a divide-and-conquer principle. The AND method discover class consistent neighbourhoods anchored to individual training samples (*divide*) and aggregate the local inter-sample class relationships with such neighbourhoods (*conquer*) for more reliably extracting the latent discrimination information during model training. Specially, to enhance the neighbourhood "purity" (class consistency), a progressive discovery curriculum for incrementally deriving more accurate neighbourhood supervision is proposed. Furthermore, to improve the model's invariance to intra-class image variations while maintaining class consistency within each neighbourhood, a *Progressive Affinity Diffusion* (PAD) is introduced to perform model-matureness-adaptive data group inference in training for more reliably revealing the underlying sample-to-class memberships. This is achieved by progressively self-discovering strongly connected subgraphs on a neighbourhood affinity graph via faithful affinity diffusion and formulating the group structure aware objective loss function.

1.3.2 Discovering Knowledge from Global Data Structure

This thesis discusses unsupervised object recognition without any human annotations in discovering knowledge from global data structure. Considering the similar objective shared with the conventional cluster analysis but is discussed in the context of deep learning, such a technique is widely known as *deep clustering* (Hershey et al. [72]; Xie et al. [200]). Deep clustering models take unlabelled images as inputs to predict their potential membership to a desired number of clusters. It is similar in some aspects to the discriminative representation learning studied in Chapter 3. However, representation learning doesn't require explicit predictions of sample's global class memberships and holds no assumption on the number of semantic classes, hence, is more suitable for instance-level objects recognition *vs*. category-level recognition (Szeliski [170]) in deep clustering. For learning visual representations which are encoding the high-level abstract information of visual data beyond the trivial low-level visual patterns, one straightforward solution is to learn by recognising the semantic class memberships of data, *i.e*. cluster analysis. However, simply learning visual representations according to clustering objectives is prone to collapse, *e.g*. a global optimum of training deep models by the objective of k-means (Lloyd [110]) to maximise the sample density within each cluster can easily be achieved by a degenerated solution which maps arbitrary samples to a single tight cluster (Yang et al. [206]). Therefore, given the massive increase of visual data available on the Internet, how to leverage them without exhaustive label annotation for learning high-level visual semantics (sample's membership to semantic classes) remains a challenging problem.

Challenges. The *first* challenge to the problem is the mutual dependences between the objectives of representation learning and cluster analysis. One commonly adopted strategy in deep clustering is to iteratively estimate cluster assignment and/or inter-sample relationships which are then used as hypotheses in supervising the learning of deep neural networks (Wu et al. [193]; Chang et al. [21, 20]; Yang et al. [207]). In ideal cases, such alternation-learning methods can approach the performance of supervised models not the least benefiting from their robustness against noisy labels. Nevertheless, the class discriminative feature representations are the prerequisites for obtaining reliable hypothetical cluster assignments using conventional clustering techniques, while reliable pseudo class labels is also the key to learning class-sensitive visual features. Hence, those approaches are susceptible to error-propagation as the process of alternating cluster assignment and representation learning is staged between two learning objectives and

any error in global clusters is accumulated during this alternation. The *second* challenge is that the clusters derived by the models trained without manual labels as constraints are usually semantically less plausible, *i.e*. there are no one-to-one mappings between the yielded clusters and the manually defined semantic classes (common visual characteristics shared within classes are assumed) can be found. This is especially challenging to another commonly adopted paradigm in deep clustering, which simultaneously learn both the representations and cluster assignments without explicit stages of alternation but by certain proxy learning objectives (Ji et al. [85, 84]). Although they can avoid mostly the problem of error-propagation, they suffer from ambiguous learning constraints due to vague connection between the training supervision and clustering objective to separate data by their visual similarity.

Solutions. *Firstly*, this thesis proposes a deep clustering method called *PartItion Confidence mAximisation* (PICA) based on the observation that inherent visual similarity among samples are implicitly encoded in their feature representations by deep neural network even if they are given different class labels in supervised learning of object recognition (Wu et al. [195]). With this in mind, although a set of visual data can be separated in numerous ways according to various criteria, assigning samples from the same semantic categories, which are likely sharing a high proportion of visual information, to different clusters will reduce the resulted within-cluster compactness and between-cluster diversity and lead to lower partition confidence. Based on this insight, PICA is designed specifically to encourage the model to learn the most *confident* clusters from all the possible solutions in order to find the most semantically plausible inter-class separation. Specifically, a partition uncertainty index is proposed to quantify how confidently a deep model can make sense and separate a set of target images when performing both feature representation learning and cluster assignment concurrently. To fit the standard mini-batch based model learning, a stochastic approximation of the partition uncertainty index is introduced to enable deep clustering with any off-the-shelf networks. *Secondly*, a novel *Semantic Contrastive Learning* (SCL) is introduced in this thesis to alleviate the impacts of unreliable pseudo labels to representation learning by the nature of contrastive learning (Wu et al. [195]; He et al. [69]) which is able to derive sample-specific visual features in a class-agnostic manner. Meanwhile, cluster structures are explicitly imposed to unlabelled training data to encourage learning a "cluster-aware" instance discriminative feature space that promotes separation of decision boundaries between clusters, leading to a plausible interpretation of the underlying semantic concepts. Dif-

ferent from the contemporary instance contrastive learning based clustering methods (Van Gansbeke et al. [181]; Li et al. [106]; Tao et al. [175]) which pull away each instance from all other samples in the feature space, SCL only pulls it away from its pseudo-negative samples in other clusters. By sharing a common contrastive negative set for all the instances in a cluster, SCL indirectly pushes them closer regardless of any intra-cluster visual dissimilarity. This enhances model's robustness to the inevitable errors from the hypothetical class memberships estimated based on the updating features. More importantly, it resolves the contradiction in the instance contrastive learning and clustering objectives (dispersing all instances *vs*. tightening each clusters) which is neglected by the recent developments on deep clustering, ensuring a discriminative representational space is learned to encode high-level visual semantics of images.

1.3.3 Transferring Knowledge from Relevant Labels

This thesis considers unsupervised object recognition with existing knowledge acquired in relevant labelled domains when studying knowledge transfer from relevant labels to unlabelled data. Despite the remarkable progress advanced by deep learning on computer vision (LeCun et al. [98]; Krizhevsky et al. [93]; He et al. [68]), the i.i.d. assumption widely held by most of the deep learning models restricts their usability in novel target domains without additional labelled training data. In general, realistic data in novel domains usually have different and unknown distributions. It is both labour-intensive to manually annotate sufficient data and computationally expensive to re-train a model in every new domain. Inspired by human cognition which always leverages the knowledge accumulated in different domains on different tasks to understand and address new tasks with unseen data, it is vital to empower machines the same ability to transfer knowledge across tasks and domains instead of learning from scratch when there are no labels available for the data or task of interest.

Challenges. The *first* challenge to the problem is the diverse and independent efforts made on a similar objective to transfer knowledge but with different assumptions on data and/or tasks, which limits their usability in different scenarios as well as their joint contributions to the field. Recent efforts on unsupervised transfer learning have different manifestations based on their assumptions on the training data, unlabelled *vs*. sparsely (incompletely) labelled, seen *vs*. unseen. For examples, unsupervised domain adaptation (Ganin and Lempitsky [49]; Tzeng et al. [179]) deals with *unlabelled* data in a target domain from the same *seen* classes in the source domain, *i.e*.

assuming an identical task in different domains. Few-shot Learning (FSL) (Ravi and Larochelle [147]; Ren et al. [149]) assumes novel categories are *seen* in limited labelled data and Zero-shot Learning (ZSL) (Ye and Guo [209]; Xian et al. [196]) assumes the target tasks are known in relation to the seen classes in the source tasks in a word-vector attribute space. However, it is impractical to assume all novel classes are seen before, or known as having exhaustive ontology in text. Therefore, it remains an open question as how to best generalise a trained model to data from *universally unseen classes* which are without known relations in text to the seen classes in source tasks. The *second* challenge to the problem is the unreliable constraints applied to the target data which are inferred according to the non-transferable knowledge in source domains. Due to the existence of domain shift (inconsistency in data distributions (Pan and Yang [134])) and discrepancy of classes, estimating semantic class memberships of target samples by the models trained in source domains is inevitably error-prone, which tends to mislead the learning of discriminative representations of unlabelled target data. However, such differences between data domains are usually unknown, thus, it is non-trivial to determine whether the constraints constructed according to the prior-knowledge acquired in source domains are transferable to and beneficial for model learning in target domains.

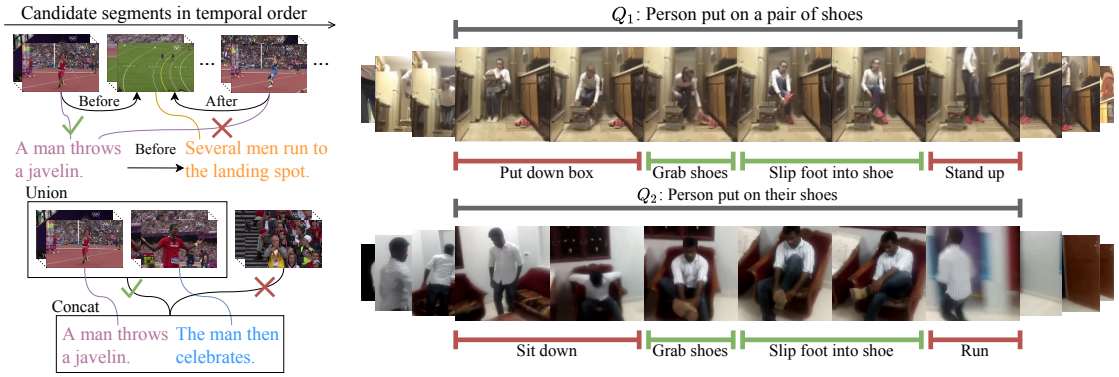
Solutions. In spite of the different assumptions made on different unsupervised transfer learning techniques, they share a common objective to map unlabelled data from novel classes into a discriminative representational space. This is identical to the long-standing clustering task while human labelling on related source data is available. Therefore, to improve model's scalability in different scenarios when knowledge transfer is needed, this thesis concentrates on such a technique named transfer clustering (Han et al. [65]), to be more generic than the above-mentioned counterparts by making no assumption on the novel categories to be seen or known. Transfer clustering models aim to predict the potential memberships of unlabelled images to a number of clusters. Their difference to the deep clustering approaches discussed in Chapter 4 is on the availability of labelled training data in a domain other than the target one. A novel method called *self-SUPervised REMEdy* (SUPREME) is introduced to that end. The motivation is that the supervision constructed according to the pre-learned knowledge acquired in a relevant source domain (transferred supervision) is not always sufficient and reliable for each and every target samples due to distribution shift and class discrepancy. Therefore, additional complementary supervision is necessary for learning a model to better describe the target distribution. Specifically, to com-

plement the unreliable pairwise relationships of target data estimated in the latent feature space yielded by the model pretrained in the source domains, the SUPREME model constructs a self-supervision to enforce consistent cluster assignment predictions of target data and their perturbed copies so to encourage model's invariance to semantic-agnostic visual distortion. By exploring the auxiliary learning principle (Caruana [19]), SUPREME explores jointly the transferred supervision acquired from human labels in relevant domains and self-supervision that is intrinsically available in the unlabelled target domains. The two supervisions are adapted on each individual target samples according to the estimated confidence of the transferred knowledge. Our empirical studies demonstrate that by exploring transfer clustering with self-supervised learning in a target domain can attain compelling knowledge transfer ability comparing to contemporary FSL and ZSL methods *without* the more strict assumptions.

1.3.4 Propagating Knowledge from Incomplete Labels

This thesis studies knowledge propagation from incomplete labels in video activity localisation by natural language (Gao et al. [50]; Zhang et al. [216]) (*a.k.a.* video moment retrieval (Mithun et al. [121]) and video grounding (Chen et al. [25])). The task aims to identify the temporal boundary of a video moment in an untrimmed and usually unscripted video according to a natural language query. A video-sentence pair is fed into the models and the potential start and end time indice of the moment described by the query sentence in the given video are predicted. The exhaustive sample-wise labels specific to such a task is intrinsically ambiguous and unaffordable regarding the labour cost which requires ones to go through the untrimmed videos frame-by-frame in order to identify the precise frame indices as the temporal endpoints of the moments-of-interest, therefore, it is of practical significance to investigate if the desired video-sentence correlations can be propagated from certain labels that may be incomplete but are easier to be collected, *e.g.* associating sentences with videos rather than video segments. Video activity localisation is fundamentally challenging as not only the visual and textual semantics should be properly understood by the models, but their matching relationships is of more importance in order to accurately locate the temporal boundary.

Challenges. The *first* challenge when learning video-text alignment from collective labels on associating video with sentences without precise temporal boundaries is that contemporary solutions locate different moments in the same videos individually, which is not optimal as it ne-



(a) Out-of-context ambiguity (b) Uncertainty in labelling activity temporal endpoints

Figure 1.2: An illustration of challenges to knowledge propagation in video activity localisation. **(a)** Video activities can be visually ambiguous if considered out of the description paragraph's context. **(b)** Activity's temporal boundaries are intrinsically uncertain in manual labelling (grey bars) which are break-down and highlighted in 'red'.

glects the fact that an individual sentence is sometimes ambiguous out of its paragraph context. For example in Figure 1.2 (a)'s top, without consideration of the *temporal relations* with the second sentence, the first query sentence (purple) can be easily mismatched with incorrect video segment, which is visually indiscriminative from the ground-truth moment. Our analysis on the ActivityNet-Captions (Krishna et al. [91]) shows that the temporal relations of over 65% moment pairs predicted by a contemporary weakly-supervised model (Lin et al. [107]) are contradictory with the true order of their descriptions. Moreover, in the bottom part of Figure 1.2 (a), "the man" in the blue query exhibits ambiguity if its *semantic relations* with previous sentences are ignored. We also observed that more than 38% descriptions in ActivityNet-Captions contain ambiguous ways of referring to expressions, *e.g.* pronouns. The *second* challenge when learning from the exhaustive temporal boundary annotations is their intrinsic ambiguity which is prone to inconsistent visual-textual correlations that is not interpretable universally. Two video moments in similar semantics but inconsistent temporal boundary are provided in Figure 1.2 (b) as examples. When the exhaustive temporal labels are available for model training, one straightforward solution is to train a model to predict directly the start and end frame indices of a target moment (Zeng et al. [214]; Ghosh et al. [54]; Zhang et al. [217]). Such paradigm deploys directly the fixed manual activity endpoints labels for model training, implicitly assuming these labels are well-defined and ignoring their inherent uncertainties. Nevertheless, there is a considerable variation in how activities occur in unconstrained scenarios. Fitting such uncertain temporal endpoints will inevitably leads to semantically mismatched visual-textual correlations and result in poor generalisation in

test. Such a challenge is also evidenced in (Mayu Otani and Heikkilä [118]) by giving the same videos and query sentences to 5 different annotators, only 35% of their annotated activity boundary are mutually agreed (with at least 50% overlaps) on ActivityNet-Captions, highlighting the extent of activity label uncertainties in model training.

Solutions. *Firstly*, to effectively propagate knowledge about visual-textual correlations from the incomplete coarse-grained video-paragraph matching relationships to the target fine-grained alignment between video moments and sentences, a weakly-supervised method for video activity localisation called *Cross-sentence Relations Mining* (CRM) is proposed in this thesis. The key idea is to explore the cross-sentence relations in a paragraph as constraints to better interpret and match complex moment-wise temporal and semantic relations in videos. Given the one-to-one moment-sentence mappings, the inherent cross-moment relations are unknown and not straightforward to be modelled in videos but intrinsically available in the paragraph descriptions. Hence, the CRM model imposes the same cross-sentencing relations to their potentially matching video moments for more reliable proposal selections. Existing weakly-supervised counterparts (Zhang et al. [224]; Ma et al. [115]; Tan et al. [173]), which locate per-sentence queries *independently*, lack fundamentally any ability to make use of the cross-sentence relations for moment proposal selection in model training. Even though such relational information is less complete than per-sentence exhaustive temporal annotation, it is intrinsically available in the paragraph descriptions which are easier to be collected. *Secondly*, this thesis introduces *Elastic Moment Bounding* (EMB) to model explicitly the label uncertainty in the temporal boundaries of an activity moment, so that the universally interpretable knowledge about video-text alignment acquired from a majority of data can be propagated to the samples with biased temporal annotations. Instead of forcing a model to fit the manually labelled *rigid* temporal endpoints, each moment is modelled by an elastic boundary with an adaptive set of candidate endpoints. The EMB model then learns to select optimally from consistent video-text correlations among semantically similar activities. This introduces model robustness to label uncertainty and refrains the model from losing its generalisation ability when dealing with unseen videos or sentences.

1.4 Contributions

The contributions made in this thesis is summarised below:

(1) Chapter 3: A novel *Anchor Neighbourhood Discovery* method is introduced for unsuper-

vised learning of discriminative visual representations based on the idea of preserving the capability of clustering for class boundary inference whilst minimising the negative impact of class inconsistency typically encountered in clusters. It is the first attempt at exploring the concept of neighbourhood for end-to-end deep learning of feature representations without class label annotations. The AND model not only generalises the idea of sample specificity learning, but also additionally considers the originally missing sample-to-sample correlation during model learning by a novel neighbourhood supervision design. A curriculum learning algorithm is further designed to gradually perform neighbourhood discovery for maximising the class consistency of neighbourhoods, therefore, enhancing the class discrimination capability. Moreover, to further improve model's sensitivity to within-class visual variations that are common in practice, a new *Progressive Affinity Diffusion* method is formulated for model-matureness-adaptive discovery of strongly connected sub-graphs during training through affinity diffusion across adjacent neighbourhoods, which are used as a self-supervision structure for more reliable unsupervised learning conducted in an end-to-end manner.

(2) Chapter 4: This thesis proposes the idea of learning the most semantically plausible clustering solution by maximising partition confidence, which extends the classical maximal margin clustering idea (Xu et al. [203]; Cortes and Vapnik [34]) to the deep learning paradigm. The proposed idea makes no strong hypothesis on local inter-sample relations and/or cluster assignment which usually leads to error-propagation and inferior clustering solutions. As an instantiation of this idea, a novel deep clustering method, called *PartItion Confidence mAximisation* is introduced. PICA is built upon a newly introduced partition uncertainty index that is designed elegantly to quantify the global separation confidence of clusters. To enable formulating a deep learning objective loss function, a novel transformation of the partition uncertainty index is further proposed. PICA can be trained end-to-end using a single objective loss function without whistles and bells (e.g. complex multi-stage alternation and multiple loss functions) to simultaneously learn discriminative visual representations and semantically plausible cluster assignment. Besides, a new *Semantic Contrastive Learning* approach is presented to benefit from sample-wise discriminative representations derived by instance contrastive learning (Wu et al. [195]; He et al. [69]) while encouraging them to be sensitive to semantic class memberships and more importantly, re-

solving the contradiction in learning simultaneously sample specificity discrimination and clustering objectives to enforce a consensus.

(3) Chapter 5: This thesis makes the first attempt at jointly exploring related-domain prior-knowledge and novel target domain self-supervised knowledge as well as their alignment by hard samples mining in transfer clustering. This is motivated by the idea of exploiting intrinsically available self-supervision as the remedy for unreliable unsupervised transfer learning to yield more discriminative modelling of the target distributions. To that end, a novel *self-SUPervised REMEdy* method is presented for transfer clustering that enables an effective implementation of leveraging related source domain prior-knowledge whilst simultaneously mitigating the negative impact of ambiguous transferred supervision by self-supervised learning in a target domain. The empirical studies carried out in this thesis show that transfer clustering by SUPREME is able to provide compelling discrimination ability when applied to other tasks such as FSL and ZSL without the need to assume human annotations either on samples or on known semantically related classes, providing a more generic solution in both the scenarios.

(4) Chapter 6: To effectively propagate the incomplete knowledge about visual-textual correlations from the video-text holistic association labels to the moment-sentence matching relationships for locating accurately video moment of interest according to a natural language sentence when the precise temporal boundary annotations is missing during training, a new weakly-supervised method called *Cross-sentence Relations Mining* is proposed. The CRM method trains a model with both temporal and semantic cross-sentence relations, which are intrinsically available in the paragraph descriptions of activities involved in untrimmed videos, to improve per-sentence temporal boundary prediction in testing. This is the first idea to develop a model using *cross-sentence relations* in a paragraph to explicitly represent and compute *cross-moment* relations in videos, so as to alleviate the ambiguity of each individual sentence in video activity localisation. Moreover, to enhance the robustness to the inherently uncertain manual annotations on the temporal boundary of video activities, this thesis further formulates a new *Elastic Moment Bounding* method to expand the fixed manual temporal endpoints to an elastic set by reinforcing directly robust content matching as a condition to accurate endpoints predictions.

1.5 Thesis Outline

The remaining chapters of this thesis are summarised and organised as follows, with its structure being depicted in Figure 1.3.

Chapter 2 presents a comprehensive literature review of the problems discussed in this thesis, covering a wide range of techniques related to generalised unsupervised learning (Definition 2), including weakly-supervised learning, transfer learning, conventional unsupervised learning of discriminative representations for either instance-level or category-level object recognition.

Chapter 3 explores knowledge aggregation from local data structure in discriminative representation learning for instance-level object recognition. This chapter introduces two methods to improve model's robustness to unreliable adjacent relationships among samples in a latent space and also to involve more intra-class image variations in local neighbourhoods to enhance model's discrimination capacity, by progressively adapting the neighbourhoods selection and construction according to the matureness of models along the training process. Experiments are carried out in extensive benchmarks from simple digit images datasets (Netzer et al. [125]) to complex and large-scale natural objects datasets like ImageNet (Deng et al. [37]). The superior results over the state-of-the-art (SOTA) methods indicate the effectiveness of the progressive learning ideas on aggregating the knowledge about the holistic class memberships from local neighbourhoods.

Chapter 4 investigates knowledge discovery from global data structure in visual object recognition, *i.e.* deep clustering. This chapter presents two novel approaches to derive semantically plausible clusters of a collection of unlabelled data by either maximising the partition confidence of the clustering solutions inspired by the maximal margin clustering algorithms (Xu et al. [203]) or encouraging the sample-specific visual features derived by instance contrastive discrimination to be cluster sensitive so that each learned cluster is encoding a set of unique and consistent visual characteristics that potentially explaining a semantic concept. Experiments on a variety of cluster analysis benchmark datasets demonstrate the superiority of the two models on learning clusters that can be mapped to the human-defined semantic classes one-to-one.

Chapter 5 studies knowledge transfer from relevant labels in visual object recognition, *i.e.* transfer clustering. This chapter explores the idea of identifying and complementing the non-transferable knowledge acquired in the labelled source domains by the intrinsic information in the unlabelled target domains to alleviate the distractions from the unreliable transferred super-

visions. The validation of this model are carried out on a wide range of natural object imagery datasets under three different protocols including transfer clustering, FSL and ZSL, whose results demonstrate the potential of the presented SUPREME model to learn discriminative visual representation on transfer learning tasks without making stringent task-specific assumptions.

Chapter 6 discusses knowledge propagation from incomplete labels in video activity localisation. Two new video activity localisation methods are proposed to implement the ideas that accurate and universal knowledge about visual-textual correlations can be propagated from the manual annotations on video activity which are incomplete in terms of either their granularity (from coarse-grained video-paragraph to fine-grained moment-sentence relationships) or certainty (from ambiguous fixed temporal endpoints to universally interpretable boundary intervals). Extensive experiments are conducted to valid the effectiveness of the two proposed models on three video activity localisation benchmark datasets collected in-the-wild involving different common scenarios like indoor daily routines, outdoor activities and cooking tutorials.

Chapter 7 presents the conclusions, discusses the remaining problems and proposes the potential directions for future study.

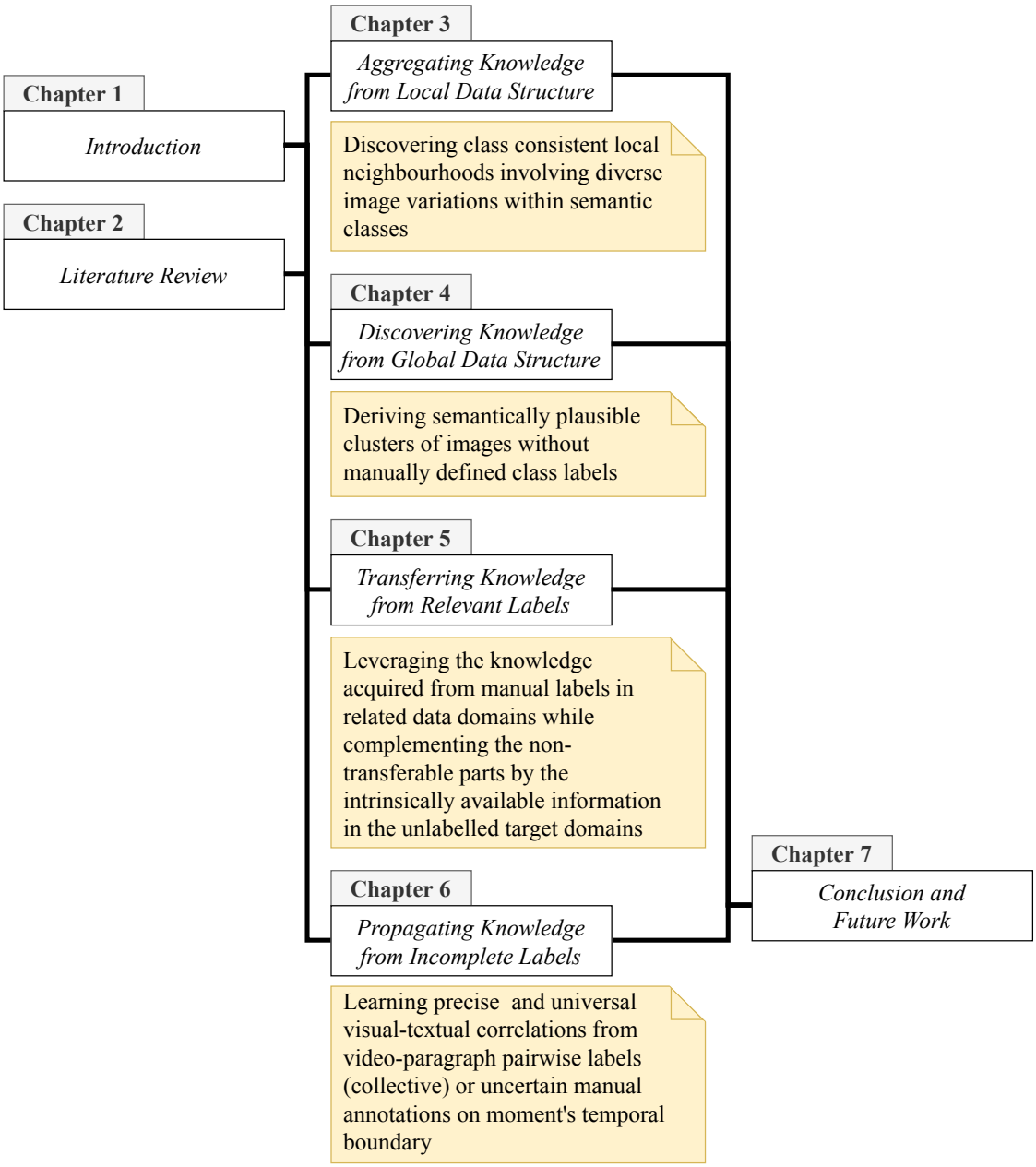


Figure 1.3: An outline of the thesis with a summary of all chapters and the structure.

Chapter 2

Literature Review

The studies carried out in this thesis are about *learning discriminative visual representations for objects or actions recognition in images or videos by deep neural networks when human annotations are insufficient and inadequate as supervisions*. This chapter provides a literature review of existing approaches which are closely related to this problem. First of all, Section 2.1 provides an overview of deep learning of visual representation in comparisons to traditional hand-crafted image descriptors; Then, representative solutions to several common and challenging problems in *generalised* unsupervised learning are discussed in Section 2.2, including knowledge discovery from data structure in representation learning (Section 2.2.1) and deep clustering (Section 2.2.2), knowledge transfer from relevant labels in transfer clustering (Section 2.2.3), and knowledge propagation from incomplete labels in video activity localisation (Section 2.2.4). Finally, a summary is given in Section 2.3.

2.1 Deep Learning of Visual Representations

Visual representation learning targets at describing and summarising the sparse visual information encoded in image pixels abstractly and holistically into a compact vector, so as to underlie the potential explanatory factors to facilitate various visual understanding tasks (Bengio et al. [12]). Desipte that substantial efforts have been made on designing delicate classification/regression algorithms (Cortes and Vapnik [34]; Ho [76]; Reynolds [150]) to push the limits of artificial visual perceptions in terms of specific applications, representation learning is always one of the most fundamental tasks by extracting meaningful information from visual data to support making rea-

sonable decisions/predictions by subsequent algorithms. Therefore, it has been widely studied both before and during the deep learning era[1].

(a) Points and patches (b) Edges (c) Straight lines

Figure 2.1: Examples of traditional image descriptors borrowed from (Szeliski [170]). Image feature detectors/descriptors in terms of (a) key points or patches (b) edges and (c) straight lines.

Traditional image feature descriptors, which are discussed out of the context of deep learning, are mostly designed heuristically by human experts to recognise certain types of visual patterns to serve as the explanatory factors for either general purposes (*e.g.* edges, textures, *etc.*) or some specific tasks (*e.g.* face recognition). Expert knowledge is the foundation of such image descriptors/features, which is usually not universal and sometimes not available. As summarised in (Szeliski [170]) and shown in Figure 2.1, traditional image descriptors (*e.g.* SIFT (Lowe et al. [113]), SURF (Bay et al. [7]), LBP (Ojala et al. [133]) and HOG (Dalal and Triggs [35]), *etc.*) are mostly formulated to describe images in terms of their **(1)** key points/patches, **(2)** edges or **(3)** straight lines and curves. The keypoint features are determined by the areas-of-interest in images, which describe the appearance of such localised areas and their surrounding pixels. For example, the well-known Harris detector (Harris and Stephens [66]) is one of the most commonly used operators to extract corners and infer features of an image. The edges descriptors, in another way, indicate the boundary of objects in images or occlusion of events in videos, which can be further considered as groups of curves and straight lines. Considering the complex visual patterns existing in natural imagery data as well as their arbitrary combinations, it is hard to decide which types of patterns can be used to bring most benefits to the target tasks. Moreover, arbitrary machine learning algorithms, like Support Vector Machines (Cortes and Vapnik [34]), Gaussian Mixture Models (Reynolds [150]), Random Forest (Ho [76]) and *etc.*, may be used as the subsequent predictors to yield the final decisions but they are usually holding different assumptions on the input features. Therefore, even though traditional image descriptors have shown remarkable

[1] In this thesis, deep learning is considered resurgent since AlexNet (Krizhevsky et al. [93]) has won the first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012.

effectiveness and efficiency on some specific computer vision applications, they still suffer from the problem of poor scalability when human expert knowledge is unavailable. Especially, take image classification as an example, it is barely possible to decide heuristically which types of patterns and their combinations can effectively reveal the difference among images of a thousand classes in ImageNet (Deng et al. [37]).

One of the most critical differences of visual feature representations derived by deep learning techniques from traditional image descriptors is that deep visual features can be learned concurrently with the target predictors in an end-to-end manner so that the essential image patterns are automatically discovered when optimising the objective of target tasks, hence, no human expert knowledge is needed in the process. In the past a decade, the emergence of deep learning (LeCun et al. [99]) has revoluted the computer vision community by its intuitive designs (mimicking the neural activity in human brain), modular formulations (composed of reusable building blocks) and remarkable capacity which surpassed human's performance in some challenging visual understanding tasks like face recognition (Taigman et al. [171]). In general, deep learning algorithms consist of **(1)** a deep neural network as the backbone model, **(2)** one or multiple learning objectives (*a.k.a.* cost function and loss function) to be optimised and **(3)** a training strategy to learn the parameters of backbone network according to the objective function. Basically, a deep neural network is a complex mapping function from the input data space of raw image pixels to the target prediction space, which can be decomposed into a set of neural layers with each independently mapping images or their features to a new latent space. Some commonly used neural layers include fully-connected layer for linear projection, convolution layer for localised linear projection with sliding windows, normalisation layers (batch normalisation (Ioffe and Szegedy [82]), instance normalisation (Ulyanov et al. [180]), *etc*.) to stablise and accelerate model learning, and a wide variety of non-linear activation functions (the family of ReLU (LeCun et al. [98]; He et al. [67]), Softmax (Goodfellow et al. [58]), *etc*.) to introduce non-linearity into the network for modelling complexer non-linear mappings. By stacking up multiple neural layers in different combinations, several types of neural networks have been shown powerful in a wide range of computer vision applications like convolutional neural networks (AlexNet (Krizhevsky et al. [93]), VGG (Simonyan and Zisserman [163]), ResNet (He et al. [68]), InceptionNet (Szegedy et al. [169]), *etc*.) in tasks related to static images, recurrent neural networks (Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber [77]), Gated

Recurrent Unit (GRU) (Cho et al. [31])) in video sequence understanding and Generative Adversarial Network (GAN) (Goodfellow et al. [57]) in image manipulation. Beyond the structure of backbone model, the learning objective function is another crucial factor to the success of deep neural networks, which implies how should the model's parameters be updated. The formulation of cost function is task-oriented, anyhow, they are usually designed to quantify the difference of model's predictions and the expected outputs which are sometimes called the ground-truth labels. For example, the cross-entropy of a predicted and a manually annotated probability distribution is usually adopted in image classification while least absolute deviations (L_1) and least square errors (L_2) can be used for regression tasks. Therefore, exhaustive and precise manual labels is usually required for training deep neural network by providing reliable supervision signals. This results in the strong dependencies of deep learning on human annotations, and also restricts its developments when seeking for more generalisable and scalable models from large-scale data.

2.2 Unsupervised Deep Learning

Unsupervised learning has been widely studied in the machine learning and computer vision communities, which aims to learn the desired mapping functions from data observations to a target numerical (regression) or discrete (classification) label space without using any human annotations. It plays a significant role in lots of computer vision applications when human annotation is unaffordable (*e.g.* medical image analysis (Shen et al. [160])). In recent years, along with the rapid development of deep learning, unsupervised learning is attracting explosively increasing attention even though it has already been studied for decades. This is because deep learning is well-known as a data-driven technique which relies heavily on both the large scale of training data and the reliability of supervision signals. However, assuming sufficient and precise manual labels are always available in any data domains for arbitrary tasks is not practical considering that annotating data is not only labour-intensive but sometimes prone to subjective annotation bias. This thesis carries out comprehensive studies of deep learning with insufficient and inadequate human annotations beyond the "pure" unsupervised learning setup to consider also weakly-supervised learning (Zhou [230]) when the manual labels are incomplete as well as transfer learning (Torrey and Shavlik [178]) when the labels are annotated on a relevant rather than the target data domain. For each of these techniques, we investigate it in visual representation learning for a concrete and representative computer vision task/scenario, including discovering knowledge from data

structure in representation learning and deep clustering, transferring knowledge from relevant labels in transfer clustering, and propagating knowledge from incomplete labels in video activity localisation.

2.2.1 Knowledge Mining in Unsupervised Representation Learning

Different from the tasks discussed in other sections which aim at learning the task-specific predictions (*e.g.* cluster assignments in deep/transfer clustering and temporal boundary in video activity localisation), unsupervised representation learning models produce image's feature as the only output, which is evaluated by constructing extra classifiers/predictors in the yielded feature space. Existing solutions of this problem generally fall into four different categories which are reviewed in this section: **(1)** Clustering analysis (Caron et al. [15]), **(2)** Sample specificity learning (Wu et al. [195]), **(3)** Self-supervised learning (Zhang et al. [219]) and Generative models (Goodfellow et al. [57]).

(1) Cluster analysis. Clustering is a long-standing approach to unsupervised machine learning (Coates and Ng [32]). With the surge of deep learning techniques, recent studies have attempted to learn discriminative visual representations by pseudo labels generated by independent clustering algorithms (Caron et al. [15, 16, 17]; Asano et al. [4]). For example, potential unlabelled data partitions can be constructed by k-means (Caron et al. [15]), Sinkhorn-Knopp algorithm (Asano et al. [4]; Caron et al. [16]) or with certain contraints on sample's assignment distributions like student's t-distribution (Xie et al. [200]) or a centering and sharpening constraint (Caron et al. [17]). Regardless, the key problem of these approaches remains the discovery of multiple class consistent clusters (or groups) on the entire training data. This is a difficult task with the complexity and solution space exponentially proportional to both the data and cluster size. It is particularly so for clustering based on under-trained feature representations, which potentially leads to class inconsistent pseudo data partitions. Therefore, these approaches tend to adopt an 'over-clustering' strategy to alleviate the misleading impacts of mixing the inherently negative samples in the same clusters, *e.g.* DeepCluster (Caron et al. [15]) and DINO (Caron et al. [17]) allocates a set of unlabelled data from 1,000 manually defined classes into 10,000 and 65,534 groups, respectively.

(2) Sample specificity learning. Sample specificity learning, *a.k.a.* instance contrastive learning, goes to the other extreme by considering every single sample as an independent class (Wu

et al. [195]; Bojanowski and Joulin [13]; He et al. [69]; Chen et al. [29, 27]). The key idea is that supervised deep learning of neural networks automatically reveals the visual similarity correlation between different classes from end-to-end optimisation. In concrete, existing contrastive learning approaches adopted a similar formulation to identify each instances and its synthesised positive counterparts from a set of negative (contrastive) samples while the key differences between different approaches are on the construction of negative sets and the representation of the positive counterparts. Given an image $\boldsymbol{I}$ and its perturbed copy $\boldsymbol{I}'$ generated by a random set of visual transformations on $\boldsymbol{I}$, the general formulation of instance contrastive learning is:

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} p; \\ p &= \frac{\cos(\boldsymbol{q},\boldsymbol{q}')}{\cos(\boldsymbol{q},\boldsymbol{q}') + \sum_{\boldsymbol{k}\in\mathcal{N}(\boldsymbol{I})}\cos(\boldsymbol{q},\boldsymbol{k})}, \ \boldsymbol{q} = f_{\boldsymbol{\theta}}(\boldsymbol{I}), \ \boldsymbol{q}' = f_{\boldsymbol{\theta}'}(\boldsymbol{I}') \end{aligned} \tag{2.1}$$

In Eq. 2.1, $\boldsymbol{q}$ and $\boldsymbol{q}'$ are the representations of the raw image $\boldsymbol{I}$ and its perturbed copy $\boldsymbol{I}'$ in respectively yielded by $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}'}$ while $\boldsymbol{k}$ is the feature of the samples in the contrastive set $\mathcal{N}$ of $\boldsymbol{I}$. The probability p implies how confident can the model identify every individual sample and the learning objective is to maximise p. For instances, the vanilla model (Wu et al. [195]) takes the whole dataset beside the target $\boldsymbol{I}$ as its negative contrastive set $\mathcal{N}(\boldsymbol{I})$ while the representation $\boldsymbol{q}'$ of the synthesised positive samples are obtained by the stale copies of $\boldsymbol{q}$ updated by momentum along the training process. MoCo (He et al. [69]; Chen et al. [29]) maintains a first-in-first-out (FIFO) queue in *fixed size* to keep track of a certain number of latest samples as the contrastive set of $\boldsymbol{I}$ while computing the perturbed representations $\boldsymbol{q}'$ by an independent network $f_{\boldsymbol{\theta}'}$ with identical structure as $f_{\boldsymbol{\theta}}$ and the parameters $\boldsymbol{\theta}'$ are updated according to $\boldsymbol{\theta}$. SimCLR (Chen et al. [27]) extracts both the features $\boldsymbol{q}$ and $\boldsymbol{q}'$ by $f_{\boldsymbol{\theta}}$ (*i.e.*, $\boldsymbol{\theta} = \boldsymbol{\theta}'$) and the negative set $\mathcal{N}$ is constructed by the samples in the same mini-batches, therefore, large mini-batches are required. Regardless the different model designs, this sort of supervision does not explicitly model the class decision boundaries as clustering analysis. Hence, it is likely to yield more ambiguous class structures and less discriminative feature representations.

(3) Self-supervised learning and Generative Model. Self-supervised learning has recently gained increasing research efforts (Doersch et al. [40]; Zhang et al. [218]; Noroozi and Favaro [131]; Noroozi et al. [132]; Zhang et al. [219]). Existing methods vary essentially in the design of unsupervised auxiliary supervision. Typically, such auxiliary supervision is hand-crafted to exploit some information intrinsically available in the unlabelled training data, such as spatial context (Doersch et al. [40]; Noroozi and Favaro [131]), spatio-temporal continuity (Wang and

Gupta [188]; Wang et al. [189]), and colour patterns (Zhang et al. [218]; Larsson et al. [96]). Due to the weak correlation with the underlying class targets, such learning methods mostly yield less discriminative models than clustering analysis. How to design more target related auxiliary supervision remains an open problem. On the other hand, generative model formulation is a principled way of learning the true data distribution of the training set in an unsupervised manner. The most commonly used and efficient generative models include Restricted Boltzmann Machines (Lee et al. [100]; Hinton et al. [75]; Tang et al. [174]), Auto-Encoders (Ng [126]; Vincent et al. [184]), and Generative Adversarial Networks (Radford et al. [143]; Goodfellow et al. [57]).

To learn class-discriminative visual representation without any manual labels, this thesis presents a novel idea of progressive neighbourhood mining in Chapter 3 with two instantiations called *Anchor Neighbourhood Discovery* and *Progressive Affinity Diffusion*. The progressive neighbourhood mining strategy replaces the clustering operation with local neighbourhood identification in a *divide-and-conquer* principle. This enables the control and mitigation of the clustering errors and their negative propagation, potentially yielding more accurate inference of latent class decision boundaries. Moreover, it introduces explicitly inter-sample relationships to encourage the feature representations to be aware of sample's class memberships, which is inherently missing in sample specificity learning (instance contrastive learning). Although the proposed AND and PAD models does not belong to the family of generative models, they potentially produce complementary feature representations due to a distinct (discriminative) modelling strategy.

2.2.2 Knowledge Discovery in Deep Clustering

As one of the most fundamental problems in machine learning, clustering (Coates and Ng [32]; Rokach and Maimon [152]) has been widely studied for decades in a wide range of computer vision fields (Achanta and Susstrunk [1]; Joulin et al. [86]). The objective of clustering is to split a set of unlabelled data into a number of subgroups (clusters) so that samples of the same subgroups are more similar than that of the different subgroups in terms of the chosen measure of similarity (Xu and Wunsch [204]). Whilst an enormous number of clustering algorithms have been devised during the past few decades, there are a few prominent categorises: **(1)** Hierarchical clustering (Nielsen [128]; Kaufman and Rousseeuw [87]); **(2)** Centroid-based methods (*e.g.* kmeans (Lloyd [110])); **(3)** Graph-based methods (*e.g.* spectral clustering (Ng et al. [127]; Shi

and Malik [161])) and **(4)** Density-based methods (*e.g*. DBSCAN (Ester et al. [46])). Regardless that the types of clustering algorithms are far more than just the few listed here, most of them are conducted on a fixed feature representational space. Thereby, the quality of the resulted clusters is restricted by that of the imagery features.

Recently, along with the rapid development of deep learning, clustering has drawn growing attention to be studied jointly with representation learning, forming a new and more challenging task called Deep Clustering (Hershey et al. [72]; Xie et al. [200]). Deep Clustering aims to separate a set of unlabelled raw data (rather than their pre-fixed representations) into a desired number of clusters by exploring their inherent global data structure. This section provide literature reviews of deep clustering techniques according to the two common training strategies adopted by them including **(1)** alternate training and **(2)** simultaneously training, which are closely related to the SCL and PICA models proposed in Chapter 4, respectively.

(1) Alternate training strategy. Alternation-learning methods (Yang et al. [206]; Xie et al. [200]; Chang et al. [20]; Wu et al. [193]; Chang et al. [21]; Yang et al. [207]; Guo et al. [62]) is one of the most straightforward solutions for deep clustering, which usually estimates the ground-truth membership according to the pretrained or up-to-date model and in return supervises the network training by the estimated information (pseudo labels). DEC (Xie et al. [200]) initialises cluster centroids by conducting k-means (Lloyd [110]) on pretrained image features and then fine-tunes the model to learn from the confident cluster assignments to sharpen the resulted predictions by encouraging one-hot distributions. IDEC (Guo et al. [62]) shares a similar spirit and improves by a local structure preservation mechanism. JULE (Yang et al. [207]) combines the hierarchical agglomerative clustering idea with deep learning by a recurrent framework which merges the clusters that are close to each other. The method in (Yang et al. [206]) jointly optimises the objectives of Auto-Encoder (Bengio et al. [10]) and k-means (Lloyd [110]) and alternately estimates cluster assignment to learn a "clustering-friendly" latent space. DAC (Chang et al. [20]), DDC (Chang et al. [21]) and DCCM (Wu et al. [193]) exploit the inter-samples relations according to the pairwise distance between the latest sample features and train the model accordingly. In idea cases, such alternate training strategy can approach the performance of supervised models by benefitting from explicit supervised discriminative learning. However, these approaches suffer from the problem of severe error-propagation from the inconsistent estimations of cluster assignment to visual feature learning during the training process.

To improve the robustness of alternate learning strategy while maintaining its discrimination ability from mimicking supervised learning, this thesis presented a novel *Semantic Contrastive Learning* model in Chapter 4. Inspired by sample-specific learning (Wu et al. [195]; He et al. [69]) which is able to discover the intrinsic visual similarity among imagery samples despite a lack of knowledge of their true class memberships, the SCL model is more robust to error propagation from the intermediate cluster assignments because it may degenerate to conventional instance contrastive learning which is still able to derive instance-wise visual features even when the estimated cluster assignments are completely random. With such a design, the SCL model is able to refrain from incorrectness accumulation during the alternation of learning representations and decision boundaries.

(2) Simultaneous training strategy. To learn data partitions without alternation, simultaneous training strategy often integrates deep representation learning (Bengio et al. [10]; Vincent et al. [184]) with conventional cluster analysis (Lloyd [110]; Elhamifar and Vidal [45]; Gowda and Krishna [60]) or other pretext objectives. These approaches do not explicitly learn from estimated cluster assignment but usually require good cluster structure to fulfil the training objectives. Methods in (Ji et al. [84]; Peng et al. [137, 138]) train to optimise the objective of cluster analysis and concurrently adopt the reconstruction constraint to avoid trivial solutions. ADC (Haeusser et al. [63]) formulates the optimisation objective to encourage consistent association cycles among cluster centroids and sample embeddings, while IIC (Ji et al. [85]) trains a model to maximise the mutual information between the predictions of positive sample pairs. Both of them randomly perturb the data distribution as a cue of positive relationships to enourage model's invariance to trivial visual deformations which are usually irrelevant to image semantics. Although the methods from this category alleviate the negative impact of inaccurate supervision from estimated information, their objectives are usually more ambiguous than that of the alternate approaches as they can be met by multiple different separations. Due to vague connection between the training supervision and clustering objective, this type of approaches tend to yield semantically less plausible clusters solutions.

Considering the ill-defined nature of deep clustering that a set of unlabelled images can intrinsically be grouped in multiple ways according to different criteria, this thesis makes an assumption on the consistency of visual and semantic similarity among image data regardless of the existence of within-class visual variation and inter-class affinity. Based on the observations that

the intrinsic visual similarity among images will be implicitly encoded by deep neural network in feature distances regardless of the assigned class labels, a new PartItion Confidence mAximisation model is proposed in Chapter 4 which introduce a partition uncertainty index to quantify the global confidence of clustering solutions so as to learn the most separable clusters which are potentially the most semantically plausible solution. The PICA method can be trained in a concise and simultaneous manner without any ad-hoc strategy for pretraining or alternation.

2.2.3 Knowledge Transfer in Transfer Clustering

Transfer Learning (Pan and Yang [134]; Torrey and Shavlik [178]) is one of the most critical areas studied in the computer vision community, aiming to levearge the knowledge encoded in human-labelled data to help understanding/modelling any new and unseen data without exhaustive labels. According to the conditions of different application scenarios, the studies of transfer learning are divided into multiple branches with various assumptions on the relationships of the labelled source data and the unlabelled target data in terms of their label spaces and data distributions. This section reviews transfer learning **(1)** first in a global picture to elaborate the different lines of works with distinct assumptions and then **(2)** investigates transfer clustering (Han et al. [65]) which is a more general task studied in this thesis and is closely related to the SUPREME model proposed in Chapter 5.

(1) Unsupervised Transfer Learning. Unsupervised Domain Adaptation (UDA) has been widely studied to transfer knowledge across different data distributions. Critically, by assuming all the domains are sharing the same label space, UDA learns to either align feature distributions (Ganin and Lempitsky [49]; Tzeng et al. [179]) or map the relationships between the decision boundaries and feature representations so that the boundaries are valid for both domains (Lee et al. [102]; Saito et al. [156]). However, UDA is not always practical as it cannot enumerate all the categories for model training let alone exhaustively collecting and annotating the data. Alternatively, Zero-shot Learning and Few-shot Learning have drawn increasing attention, in which target data of novel classes are *unseen* or weakly-seen during model training. As generalising CNN models to *universal unseen* categories is intrinsically challenging, both these two tasks hold an assumption that the novel classes are *known* according to their pre-defined relationships with the seen classes. In ZSL (Song et al. [164]; Chao et al. [23]), all the seen (source) and unseen (target) classes are described by a common representational space, *e.g.* attribute space, so that

data from unseen classes can be classified by novel combinations of attributes which are building blocks shared with the seen classes. Different from ZSL, FSL (Ye et al. [208]; Miller et al. [119]) assumes that a few (usually $1 \sim 5$) anchor samples with manual labels are available from each novel classes as its prototype (so strictly not unseen), and the novel classes and seen classes are in proximity in their distributions. In this case, the target data can be classified according to their distances to different anchor samples with the assistance of the distributions of seen classes.

Considering the common objectives of UDA/FSL/ZSL to derive discriminative feature representations of the unlabelled target data with the help of the human annotations on relevant data, their individual assumptions limit their deployability and usability in different real-world tasks. To address such a limitation, the proposed SUPREME method aims to deal with *universal unseen* categories without human annotation in any level, neither semantically nor visually. This is intrinsically more challenging than UDA/FSL/ZSL by holding fewer assumptions but is more generic in practice and is potentially beneficial to all the above-mentioned tasks.

(2) Transfer clustering. To partition unlabelled target data with pre-learned source domain knowledge, Han *et al*. (Han et al. [64]) first introduced the task of transfer clustering formally, which aims to jointly learn *both* the representations and the decision boundaries of unlabelled target data with the help of the labelled data from related domains. They construct the initial clustering solution by applying k-means upon the feature representation produced by a pretrained model and learn to sharpen the initial assignment distribution of each sample. As the follow-up, Han *et al*. (Han et al. [65]) propose to learn the intial feature space by self-supersion and transfer knowledge by rank statistics. They train the model by the confident supervision constructed according to prior-knowledge and learns to gradually sharpen the initial prediction distribution of each sample. There are other attempts at transferring knowledge across domains for achieving learning tasks that are similar to transfer clustering. For example, KCL (Hsu et al. [79]) and MCL (Hsu et al. [80]) are based on a Constrained Clustering Network, which learns to transfer pairwise similarities from a source to a target domain so that the cross-domain and cross-task transfer learning are decoupled. This aims to reduce the model learning complexity. Another method called Centroid Networks (Huang et al. [81]) is proposed to jointly learn data embeddings and clustering using the Sinkhorn k-means algorithm. Due to distribution shift and discrepancy in label space proximity, the clustering of some target samples during model training will not yield sufficient supervision from the prior-knowledge of the source domain, therefore, poorly

modelled.

Whilst the non-transferable knowledge obtained in relevant labelled data is inevitably existing when being applied to the unlabelled target data, transferring such knowledge will certainly result in ambiguous and unreliable supervisions that reduce the class discrimination ability of the learned representations for the target data. Therefore, the SUPREME model presented in Chapter 5 tackles this problem by integrating self-supervision into transfer learning as the complementary of insufficient transferred supervisions so to better explain the target distribution. Although Han *et al*. (Han et al. [65]) adopted self-supervised learning for model pretraining, it is conducted as a separated step, hence, failed to exploit the complementary of the two supervisions and lead to suboptimal results.

2.2.4 Knowledge Propagation in Video Activity Localisation

The problem of lacking reliable supervisions for training deep neural networks does not only exist in pure unsupervised learning or transfer learning, but also when human annotations are incomplete. For examples, two types of incomplete labels commonly seen in practice are when they are given in collective in weakly-supervised learning (Zhou [230]; Guo et al. [61]) or annotated with ambiguity/uncertainty in fully-supervised learning (Zhou et al. [229]). To investigate effective algorithms for deep model training with such incomplete labels, video activity localisation by natural language (Gao et al. [50]; Heilbron et al. [70]) is a perfect application to be studied since the temporal boundary annotations for video activity is intrinsically ambiguous (uncertain labels) and associating natural language sentences with videos (collective labels) is much easier than identifying the precise start and end time indices of video moments. In this regard, this thesis discusses the problem of knowledge propagation from incomplete labels in video activity localisation. This section reviews: **(1)** fully-supervised localisation methods based on content alignment or boundary identification; **(2)** intrinsic uncertainty in temporal boundary and contemporary attempts on resolving it, and **(3)** existing weakly-supervised solutions to localisation without requiring exhaustive temporal boundary annotations during model training.

(1) Content Alignment and Boundary Identification. The two commonly adopted training strategies in video activity localisation (Gao et al. [50]; Heilbron et al. [70]) learn to locate by matching the content of moment proposals with query sentences in segment level or directly predict the boundary of target moments in frame level. Technically, these two types of approaches

can be identified by whether they are with a *proposal-based* or *proposal-free* design. By aggregating all the frames within a video segment (proposal) and aligning them holistically with the query sentences, the proposal-based content alignment approaches (Anne Hendricks et al. [3]; Gao et al. [50]; Zhang et al. [221, 220]; Ge et al. [52]) are insensitive to the boundary as its most salient and semantically aligned parts are not necessarily at its two ends. However, the endpoint frames usually play a significant role in differentiating video moments from their overlapping counterparts containing redundant frames, hence, critical for video activity localisation. In contrast, the proposal-free boundary identification methods (Chen et al. [25]; Zeng et al. [214]; Ghosh et al. [54]; Zhang et al. [217]; Nan et al. [124]; Zhou et al. [229]; Zhao et al. [228]; Li et al. [105]) learn to directly regress the start and end timestamps of the target moments or predict the per-frame probabilities of being the endpoints. In either case, they take the temporal boundaries provided manually as the oracles for learning exactly the same predictions. However, this is prone to be misled by the uncertainty in manual labels and results in less generalisable models. What's more, there are some recent attempts at benefitting from both the learning strategies, which are called *Joint content-boundary learning* models in this thesis. They mostly adopted a dual-branch designs to encode video in frame- and segment-level independently, then explore the interaction (*e.g*. assembling) of frame's and segment's features to explore visual information in different granularities for better video comprehension.

(2) Temporal Boundary Uncertainty. The uncertainty of temporal boundary annotation refers to the difference of boundary labels annotated by different workers according to the same natural language sentences as the queries. Otani *et al*. (Mayu Otani and Heikkilä [118]) quantitatively studied this problem by collecting multiple boundaries for the same activities from different annotators. Giving the same videos and query sentences to 5 different annotators, only 42% and 35% of their annotated activity boundary are mutually agreed (with at least 50% overlaps) on Charades-STA (Gao et al. [50]) (filming indoor daily routines) and ActivityNet-Captions (Krishna et al. [91]) (mostly about outdoor activities), respectively. This highlights the extent of activity label uncertainties in model training inherent to the current proposal-free methods, and the potential significant misinformation in training such models. However, Otani *et al*. (Mayu Otani and Heikkilä [118]) did not explicitly propose a solution to the problem. DeNet (Zhou et al. [229]), on the other hand, addressed the problem of uncertainty w.r.t. the variety of language descriptions, *i.e*., the same video activity can be described semantically in different ways. They

generated different copies of the same query sentences by perturbing the "modified" phrases (adjective, adverb and *etc*.) so to predict diverse boundaries for the same video activities. Besides, there are a few other attempts at modelling uncertain boundary by Gaussian distributions (Chen et al. [30]; Xie et al. [201]; Wang et al. [186]; Xiao et al. [199]). However, labelling uncertainty is usually random and unpredictable, unlikely Gaussian in general.

To explicitly model the boundary uncertainty during training and get rid of the misinformation to derive universally interpretable visual-textual correlation from such uncertain annotations, a new *Elastic Moment Bounding* method is introduced in Chapter 6. The EMB model explores collaboratively both proposal-free and proposal-based mechanisms for learning and expand manually annotated a single pair of fixed activity endpoints to an elastic set, so as to reinforce directly robust content matching (the spirit of proposal-based) as a condition to accurate endpoints localisation (the spirit of proposal-free) of activities in videos. The presented EMB method explores the combination of both the proposal-based and proposal-free learning strategies for attention learning of activity temporal boundary conditions beyond feature learning for activity representation. It augments the fixed manual labels by the video segments selected according to their content alignments with query sentences to help improve the robustness of temporal endpoints identification when there is boundary uncertainty. Rather than studying the uncertainty from the perspective of semantic description in DeNet (Zhou et al. [229]), this thesis analyses it from the perspective of inherent uncertainties in the activity temporal boundary annotations, which is intrinsically harder to avoid.

(3) Locating Video Activity with Weak Supervision. In the absence of temporal boundary annotations, most of the existing weakly-supervised approaches are either based on Multi-Instance Learning (MIL) (Keeler et al. [88]) or jointly learn with reconstruction task. The MIL-based methods (Gao et al. [51]; Tan et al. [173]; Ma et al. [115]; Zhang et al. [224]) learn the video-text alignment in the video-level by maximising the matching scores of the videos and their corresponding descriptions manually annotated on the datasets while suppressing that of the videos and the descriptions of others. Such learned video-text alignment is then applied to locate the moments which are best matched with the given queries in inference. Another commonly adopted reconstruction strategy (Lin et al. [107]; Duan et al. [43]) aims at selecting the video segments which can help accomplish the reconstruction task to the largest extent, *e.g*. WS-DEC (Duan et al. [43]) jointly optimises the sentence localisation and video captioning tasks so to identify

the video segments which yield consistent captions with the queries.

Even though remarkable progress has been made by the abovementioned works in the past few years, none of these methods fully exploit the video-level paragraph descriptions but treat different sentences in the paragraph independently. In Chapter 6, a new *Cross-sentence Relations Mining* approach is proposed to explore the relations of sentences in paragraphs and use them to constrain the selections of moments in training so that only the reliable video segments with consistent relations will be aligned with the query sentences. Although such relational information is less complete than per-sentence exhaustive temporal annotation, it requires no extra annotations and avoids subjective bias from inherent ambiguity in temporal labelling (Mayu Otani and Heikkilä [118]).

2.3 Summary

This section provides a comprehensive review of deep visual representation learning in the perspective of training deep neural network with insufficient and inadequate human annotations. As summarised in Table 2.1, several common scenarios in challenging computer vision tasks are discussed, including discovering knowledge without any labels in representation learning and deep clustering, transferring knowledge from relevant labels in transfer clustering, and propagating knowledge from incomplete labels in video activity localisation. A wide variety of existing works have been studied, which have yielded impressive performances but still suffer from some limitations. The following chapters investigate several unresolved challenges in these problems, for which novel deep learning algorithms are introduced. Specifically:

(1) (Chapter 3) **Knowledge aggregation from local data structure in representation learning**: Among the existing solutions for learning visual representations, joint clustering approaches and sample specificity learning have been shown superior on deriving discriminative visual features. Nevertheless, considering the complexity and solution space exponentially proportional to both the data and cluster size, the clustering-based approaches are subjective to error-propagation. In contrast, sample specificity learning takes every individual instance as a pseudo class augmented by linear and global image transformations, which results in instance discriminative features that are less sensitive to the underlying semantic classes involving complex non-linear image variations. this thesis presents a progressive neighbourhood mining strategy to combines the advantages of both clustering and

sample specificity learning whilst mitigating their disadvantages by a divide-and-conquer principle to discover class consistent neighbourhoods anchored to individual training samples and propagtes the local adjacent relationships across neighbourhoods for more reliably extracting class discriminative information in training.

(2) (Chapter 4) **Knowledge discovery from global data structure in deep clustering**: Cluster analysis have been a long-standing problem in unsupervised learning, however, to be conducted simultaneously with representation learning, existing methods suffer from either error-propagation in the alternation of clusters and features learning or the ambiguous learning constraints due to vague connections between the training supervision and clustering objective which will result in semantically less plausible data separations. To resolve the problem of error-propagation, a new variant of sample-specific contrastive learning called Semantic Contrastive Learning is proposed, which derives visual features in a class-agnostic manner that is robust to unreliable intermediate cluster assignments estimated in the process of model training. Besides, by extending the classical maximal margin idea to the deep learning paradigm, this thesis introduces a novel PartItion Confidence mAximisation model to derive global decision boundary of clusters by maximising data separation confidence, so as to ensure the semantic plausibility of the learned clusters.

(3) (Chapter 5) **Knowledge transfer from relevant labels in transfer clustering**: Whilst unsupervised transfer learning have been extensively studied in several independent lines of works with their own assumptions on the relationships of source and target tasks and data distributions, it is intrinsically more challenging to conduct knowledge transfer without assuming the target data is seen or the target classes are known, which is barely investigated in existing works. To facilitate knowledge transfer in a more generic scenario without making over strong assumptions, this thesis presents a self-SUPervised REMEdy model for clustering of unlabelled image data with the help of the manual labels in a relevant data domains. The SUPREME model identifies non-transferable knowledge from source to target domains by discovering hard target samples with insufficient and ambiguous supervisions. After that, SUPREME construct self-supervisions by intrinsically available positive inter-sample relationships in the target domain to complement those weak transferred supervisions, so as to ensure the discrimination ability of the resulted visual features as well as the separability of the yielded clusters.

(4) (Chapter 6) **Knowledge propagation from incomplete labels in video activity localisation**: Existing video activity localisation approaches rely heavily on the temporal boundary annotations of video moments to derive visual-textual correlations while such exhaustive labels are not only labour-intensive to be collected but also prone to subjective labelling bias, which are likely lead to ambiguous and unreliable supervision signals for deep model training. To accurately locate video activity by natural language, a novel Cross-sentence Relations Mining method is proposed in this thesis, which explores video-text alignments according to the intrinsically available cross-sentence relations in the paragraph description of videos so that no moment-wise temporal boundary annotation is needed for model learning. Moreover, this thesis further investigates another common case of incomplete labels when the human annotations are uncertain. To this end, a new Elastic Moment Bounding approach is introduced to derive universally interpretable visual-textual correlations by expanding manually annotated a single pair of fixed activity endpoints to an elastic set and selecting optimally among semantically similar activities.

Task	Solutions	Remarks
Representation Learning	Cluster analysis	It is difficult to discover class consistent clusters because the complexity and solution space exponentially propotional to the data and cluster size.
	Sample specificity learning	Per sample specific feature representations are learned, which does not explicitly model sample's potential class memberships.
	Self-supervised learning and Generative model	Expressive visual knowledge is acquired by pretext learning task. However, how to design more target related auxiliary supervision remains an open problem.
Deep Clustering	Alternate learning	Benefiting from mimicking supervised learning but is prone to errors accumulation
	Simultaneous learning	Suffering from the vague connection between the pretext training supervision and clustering objective.
Transfer Clustering	Unsupervised transfer learning	UDA, FSL and ZSL hold different assumptions on the target data and label spaces, hence, being limited to specific scenarios
	Transfer clustering	Overlooking the hard sample with ambiguous transferred supervisions, which are important for learning discriminative feature representations of target data
Video Activity Localisation	Fully-supervised learning	Content alignment is less sensitive to accurate temporal endpoints while boundary identification is fragile as the labels are inherently uncertain
	Weakly-supervised learning	Doesn't require exhaustive temporal boundary labels for training but ignores the cross-sentences relations which are intrinsically available

Table 2.1: A summary of existing solutions to different computer vision tasks related to generalised unsupervised learning including unsupervised representation learning, deep clustering, transfer clustering and video activity localisation.

Chapter 3

Aggregating Knowledge from Local Data Structure by Neighbourhood Mining

---

This chapter discusses knowledge aggregation from local data structure in learning class discriminative visual representations without any human annotations on pre-defined semantic categories. Suppose we have N training images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_N\}$ with no sample-wise exhaustive class labels being available. The objective is to derive a deep CNN model $\boldsymbol{\theta}$ from the imagery data $\mathcal{I}$ that allows to extract class discriminative feature representations $\boldsymbol{x}$, $f_{\boldsymbol{\theta}} : \boldsymbol{I} \rightarrow \boldsymbol{x}$. Without the access to class labels, it is unsupervised how the feature points $\boldsymbol{x}$ should be distributed in training so that they can correctly represent the desired class memberships. It is therefore necessary for an unsupervised learning algorithm to reveal such discriminative information directly from the visual data. This is challenging due to the arbitrarily complex appearance patterns and variations typically exhibited in the image collections both within and across classes, implying a high complexity of class decision boundaries.

The structure of this chapter is as follows: An overview of the idea to learn from local neighbourhoods and its unique merits over the contemporary unsupervised representation learning approaches are presented in Section 3.1; Section 3.2 introduces the detailed formulation of learning discriminative feature representations by neighbourhoods discovery characterised by a curriculum learning strategies to get rid of the negative impacts of the inevitable false-positive inter-sample relationships; Section 3.3 further provides a model-matureness-adaptive design for finding a trade-off between the class consistency and visual variations within local neighbour-

hoods; Section 3.4 provides thorough experiments to valid the effectiveness of the proposed models on extensive object recognition benchmark datasets; Section 3.5 summarises the ideas to aggregate knowledge from local data structure for learning discriminative feature representations.

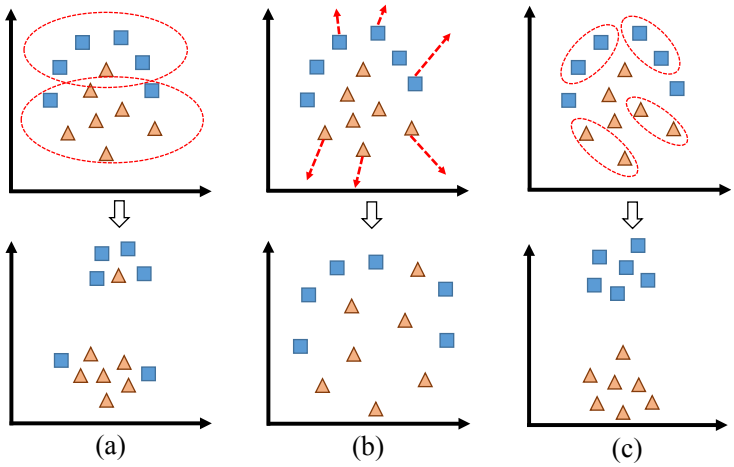


Figure 3.1: An illustration of three unsupervised representation learning strategies. **(a)** *Clustering analysis* aims for discovering the global class decision boundary (Caron et al. [15]; Xie et al. [200]); **(b)** *Sample-specificity learning* discards the concept of clusters by treating every training sample as an independent class (Wu et al. [195]; Bojanowski and Joulin [13]); **(c)** Our *Neighbourhood Discovery* searches local neighbourhoods with high class consistency.

3.1 Approach Overview

To effectively discovery semantic class memberships among samples to derive visual features that is sensitive to both intra-class discrepancy and inter-class similarity, we take a *divide-and-conquer* strategy from the local sample-anchored neighbourhood perspective. Specifically, an intuitive method to determine such neighbourhoods is using k nearest neighbours (kNN) given a feature space $\mathcal{X}$ and a similarity metric, *e.g.* the cosine similarity. A neighbourhood $\mathcal{N}_k(\boldsymbol{x})$ determined by kNN is sample-wise, *i.e.*, anchored to a specific training sample $\boldsymbol{x}$:

$$\mathcal{N}_k(\boldsymbol{x}) = \{\boldsymbol{x}_i \mid \cos(\boldsymbol{x}_i, \boldsymbol{x}) \text{ is top-}k \text{ in } \mathcal{X}\} \cup \{\boldsymbol{x}\}, \tag{3.1}$$

where $\mathcal{X}$ denotes the feature space. We call such structures as ***Anchor Neighbourhoods*** (ANs). The key idea is that, whilst it is difficult and error-prone to directly reason the *global class decision boundaries* at the absence of class labels on the training data (Figure 3.1 (a)), it would be easier and more reliable to estimate *local class relationship* in small neighbourhoods (Figure 3.1 (c)). Although such information is *incomplete* and provides *less* learning supervision than the conventional clustering strategy (Caron et al. [15]; Xie et al. [200]) that operates at the coarse

group level and mines the clusters of data samples, it favourably mitigates the misleading effect of unreliable supervision. Besides, the idea of learning from local data structure differs dramatically from the sample-specificity learning strategy (Wu et al. [195]; Bojanowski and Joulin [13]), that lacks a fundamental ability to mine the inter-sample class relationships primitive to the global class boundaries (Figure 3.1 (b)). Therefore, it represents a conceptual trade-off between the two existing strategies and a principled integration of them.

3.2 Anchor Neighbourhood Discovery

This section introduces the proposed *Anchor Neighbourhood Discovery* method for unsupervised discriminative representation learning.

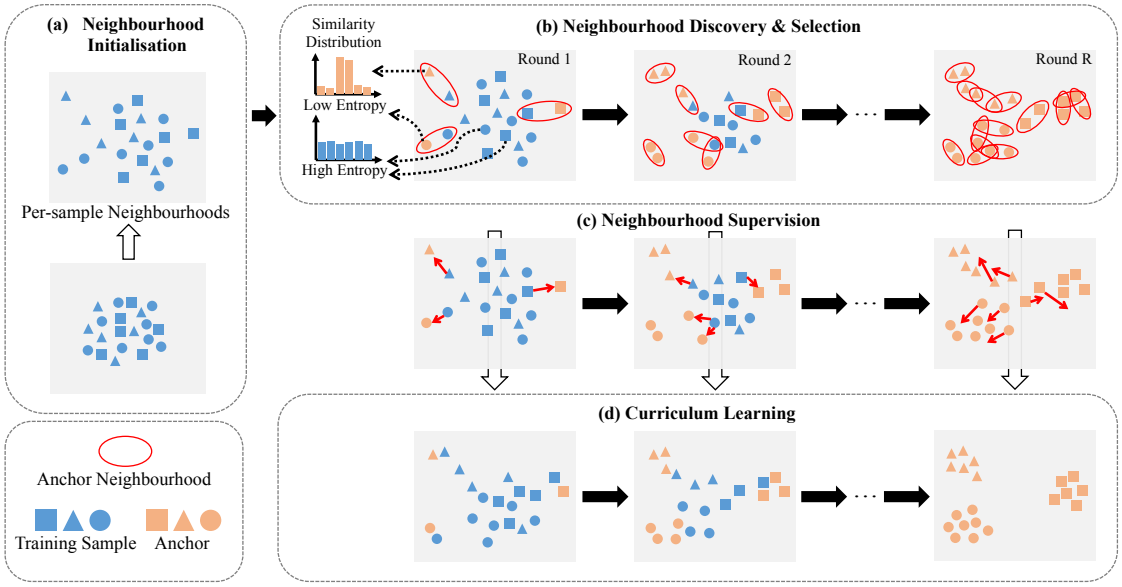


Figure 3.2: An overview of the proposed *Anchor Neighbourhood Discovery* method for unsupervised discriminative representation learning.

An overview of the introduced AND model is depicted in Figure 3.2. **(a)** The AND model starts with per-sample neighbourhoods for model initialisation. **(b)** The resulting feature representations are then used to discover the local neighbourhoods anchored to every single training sample, *i.e.*, anchor neighbourhoods. **(c)** To incorporate the neighbourhood structure information into model learning, we propose a differentiable neighbourhood supervision loss function for enabling end-to-end model optimisation. **(d)** For enhancing model discriminative learning, we further derive a curriculum learning algorithm for selecting class consistent neighbourhoods in a progressive manner. This is based on a novel similarity distribution entropy measurement.

3.2.1 Neighbourhood Discovery

To enable class discriminative learning, we want all samples in a single neighbourhood AN to share the same class label, *i.e.*, *class consistent*. As such, we can facilitate the design of learning supervision by assigning the same label to these samples. This requirement, however, is non-trivial to fulfil in unsupervised learning since we have no reasonably good sample features, even though a neighbourhood AN can be much smaller and more local (therefore likely more class consistent) than a typical cluster when using small k values. Moreover, we begin with the training images but *no* learned features. This even prevents the formation of $\mathcal{N}_k$ and gives rise to an extreme case – each individual sample represents a distinct anchor neighbourhood.

Neighbourhood Initialisation. Interestingly, such initial ANs are in a similar spirit of sample-specific instance contrastive learning (Wu et al. [195]; Bojanowski and Joulin [13]) where each data instance is assumed to represent a distinct class (Figure 3.2 (a)). With this conceptual linkage, we exploit the instance loss (Wu et al. [195]) to commence the model learning. Specifically, it is a non-parametric variant of the softmax cross-entropy loss written as:

$$\begin{aligned} \mathcal{L}_{\text{init}} &= -\sum_{i=1}^{n_{\text{bs}}} \log(p_{i,i}), \\ p_{i,j} &= \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{x}_j/\tau)}{\sum_{k=1}^{N} \exp(\boldsymbol{x}_i^\top \boldsymbol{x}_k/\tau)}, \end{aligned} \tag{3.2}$$

where $p_{i,i}$ indicates the normalised probability that sample x_i can be correctly identified from all the other instances while n_{bs} denoting the training mini-batch size and the temperature parameter τ is for controlling the distribution concentration degree (Hinton et al. [74]).

Neighbourhood Supervision. In the feature space derived by Eq. (3.2), we build a neighbourhood $\mathcal{N}_k(\boldsymbol{x}_i)$ for each individual sample $\boldsymbol{x}_i$. Considering the high appearance similarity among the samples of each $\mathcal{N}_k(\boldsymbol{x}_i)$, we assume they share a single class label for model discriminative learning. Formally, we formulate an unsupervised neighbourhood supervision signal as:

$$\mathcal{L}_{\text{AN}} = -\sum_{i=1}^{n_{\text{bs}}} \log\Big(\sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} p_{i,j} \Big) \tag{3.3}$$

The rationale behind Eq. (3.3) is to encourage label consistency for anchor neighbourhoods (Figure 3.2 (c)). Specifically, the probability $p_{i,j}$ (Eq. (3.2)), obtained using a softmax function, represents visual similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in a stochastic manner. This takes the spirit of *stochastic nearest neighbour* (Goldberger et al. [56]), as it considers the entire training set. In this scheme, the probability $P(\boldsymbol{x}_i)$ of correctly classifying a sample $\boldsymbol{x}_i$ can be then represented as:

$$P(\boldsymbol{x}_i) = \sum_{j \in C} p_{i,j} \tag{3.4}$$

where C denotes the set of samples in the same class as $\boldsymbol{x}_i$. However, C is unavailable to unsupervised learning. To overcome this problem, we approximate C by the neighbourhoods ANs, each of which is likely to be class consistent. Together with the cross-entropy function, this finally leads to the formulation of the proposed $\mathcal{L}_{\text{AN}}$ loss (Eq. (3.3)).

Remarks. The proposed neighbourhood supervision formulation $\mathcal{L}_{\text{AN}}$ (Eq. (3.3)) aims at exploring the *local class information*, under the assumption that anchor neighbourhoods are class consistent. This is because each neighbourhood AN is treated as a different learning concept (*e.g.* class), although some ANs may share the same unknown class label. Such information is also *partial* due to that a specific AN may represent only a small proportion of the corresponding class, and multiple ANs with the same underlying class can represent different aspects of the same concept *collectively* (not the whole view due to no AN-to-AN relationships). It is the set of these distributed anchor neighbourhoods *as a whole* that brings about the class discrimination capability during model training. It is in a *divide-and-conquer* principle.

Fundamentally, the proposed design differs dramatically from both (1) *the clustering strategy* that seeks for the complete class boundary information – a highly risky and error-prone process (Caron et al. [15]; Xie et al. [200]), and (2) the *sample specificity learning* that instead totally ignores the class level information therefore less discriminative (Wu et al. [195]; Bojanowski and Joulin [13]). Moreover, clustering often requires the prior knowledge of the cluster number, therefore, limiting their usability and scalability due to the lack of it in many applications. On the contrary, this kind of information is not needed for forming the proposed ANs, hence, more application generic and scalable. To maximise the class consistency degree in ANs, we simply need to use the smallest neighbourhood size, *i.e.*, $k = 1$.

3.2.2 Neighbourhood Selection

As discussed above, the proposed method requires the neighbourhoods ANs to be class consistent. This condition, nonetheless, is difficult to meet. Specifically, the instance loss function $\mathcal{L}_{\text{init}}$ (Eq. (3.2)) encourages the feature representation learning towards that each sample's specificity degree can be maximised as possible on the training data. Considering a sample $\boldsymbol{x}_i$, other samples either share the class label (positive) with $\boldsymbol{x}_i$ or not (negative). Hence, this formulation may yield

a model with certain discrimination ability, *e.g.* when a subset of (unknown) positive samples are associated with similar visual specificity. But this entirely depends on the intrinsic data properties without stable guarantee. It means that typically *not* all neighbourhoods ANs are reliable and class consistent. This inevitably leads to the necessity of conducting neighbourhood selection for more reliable model learning.

To this end, we go beyond by taking advantages of the curriculum learning idea(Bengio et al. [11]; Dong et al. [41]). Instead of taking a one-off neighbourhood selection, we introduce a *progressive* selection process (Figure 3.2 (d)) which distributes evenly the neighbourhood selection across R rounds. This realises an easy-to-hard learning procedure through a curriculum.

Selecting Curriculum. To enable automated neighbourhood selection for making a scalable curriculum, it is necessary for us to derive a selecting criterion. This is achieved by exploiting the intrinsic nature of the probability $p_{i,j}$ (Eq. (3.2)) defined between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. More specifically, we utilise the entropy measurement of the probability vector $p_i = [p_{i,1}, p_{i,2}, \cdots, p_{i,N}]$ as the class consistency indicator of the corresponding neighbourhood AN as:

$$H(\boldsymbol{x}_i) = -\sum_{j=1}^{N} p_{i,j} \log(p_{i,j}). \tag{3.5}$$

We consider that smaller $H(\boldsymbol{x}_i)$ values correspond to more consistent neighbourhoods. In particular, when $H(\boldsymbol{x}_i)$ is small, it means $\boldsymbol{x}_i$ resides in a low-density area with sparse visual similar neighbours surrounding. In the definition of sample specificity learning (Eq. (3.2)), the model training tends to converge to some local optimum that all samples of a neighbourhood $\mathcal{N}_k(\boldsymbol{x}_i)$ with small $H(\boldsymbol{x}_i)$ share some easy-to-locate visual appearance, and simultaneously the same underlying class label statistically since positive samples are more likely to present such appearance commonness including the context than negative ones. On the contrary, a large $H(\boldsymbol{x}_i)$ implies a neighbourhood $\mathcal{N}_k(\boldsymbol{x}_i)$ residing in a dense area, a case that the model fails to identify the sample specificity. This is considered hard cases, and requires more information for the model to interpret them.

In light of the observations above, we formulate a linear curriculum according to the class consistency entropy measurement. Specifically, for the r-th round (among a total of R rounds), we select the top-S (Eq. (3.6)) of ANs according to their corresponding entropy for model learning by the proposed neighbourhood supervision loss $\mathcal{L}_{\text{AN}}$ (Eq. (3.3)).

$$S = \frac{r}{R} * 100\% \tag{3.6}$$

Since the remaining training samples are still not sufficiently interpreted by the model at the current round, they are preserved as individuals (*i.e.*, single-sample neighbourhoods) as in sample specificity learning (Eq. (3.2)).

3.2.3 Model Training

With the progressive neighbourhood discovery as above, we obtain the model objective loss function for the r-th round to be minimised for learning the model weights as:

$$\mathcal{L}^r = -\sum_{i\in B^r_{\text{inst}}} \log(p_{i,i}) - \sum_{i\in B^r_{\text{AN}}} \log\Big(\sum_{j\in\mathcal{N}_k(\boldsymbol{x}_i)} p_{i,j}\Big) \tag{3.7}$$

where B^r_{inst} and B^r_{AN} denote the set of instances and the set of ANs in a mini-batch at the r-th round, respectively. The $\mathcal{N}_k(\boldsymbol{x}_i)$ defined in Eq. (3.1) is the neighbourhood anchored at $\boldsymbol{x}_i$.

As each round of training is supposed to improve the model, we update the neighbourhoods ANs for all training samples before performing neighbourhood selection per round. To facilitate this process, we maintain an offline memory to store the feature vectors. We update the memory features of mini-batch samples by exponential moving average (Lucas and Saccucci [114]) over the training iterations as:

$$\tilde{\boldsymbol{x}}_i = (1-\eta)\cdot\tilde{\boldsymbol{x}}_i + \eta\cdot\boldsymbol{x}_i \tag{3.8}$$

where η denotes the update momentum, $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_i$ the up-to-date and memory feature vector respectively.

Model Optimisation. The proposed loss function (Eq. (3.7)) is differentiable therefore enabling the stochastic gradient descent algorithm for model training. In particular, when $\boldsymbol{x}_i$ comes as an individual instance, the gradients for $\mathcal{L}^r$ w.r.t. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($j\neq i$) are written as:

$$\frac{\partial\mathcal{L}^r}{\partial\boldsymbol{x}_i} = \frac{1}{\tau}[\sum_{k=1}^{N}(p_{i,k}\cdot\boldsymbol{x}_k) + (p_{i,i}-2)\cdot\boldsymbol{x}_i],\ \ \frac{\partial\mathcal{L}^r}{\partial\boldsymbol{x}_j} = \frac{1}{\tau}p_{i,j}\cdot\boldsymbol{x}_i \tag{3.9}$$

When $\boldsymbol{x}_i$ corresponds to an AN, the gradients are then:

$$\frac{\partial\mathcal{L}^r}{\partial\boldsymbol{x}_i} = \frac{1}{\tau}[\sum_{k=1}^{N}(p_{i,k}\cdot\boldsymbol{x}_i) - \sum_{k\in\mathcal{N}_k(\boldsymbol{x}_i)}\tilde{p}_{i,k} + (p_{i,i}-\tilde{p}_{i,i})\cdot\boldsymbol{x}_i] \tag{3.10}$$

$$\frac{\partial\mathcal{L}^r}{\partial\boldsymbol{x}_j} = \begin{cases} \frac{1}{\tau}[p_{i,j}\cdot\boldsymbol{x}_i - \tilde{p}_{i,j}\cdot\boldsymbol{x}_i], & j\in\mathcal{N}_k(\boldsymbol{x}_i) \\ \frac{1}{\tau}[p_{i,j}\cdot\boldsymbol{x}_i], & j\notin\mathcal{N}_k(\boldsymbol{x}_i) \end{cases} \tag{3.11}$$

where $\tilde{p}_{i,j} = p_{i,j}/\sum_{k\in\mathcal{N}_k(\boldsymbol{x}_i)} p_{i,k}$ is the normalised distribution over the neighbours. The whole model training procedure is summarised in Algorithm 1.

Algorithm 1: AND for discriminative representation learning.

Input: Training data $\mathcal{I}$, rounds N_{rd}, iterations per round N_{it}.

Output: A class discriminative CNN feature representation model.

```
// Initialisation
```

Model initialisation by instance specificity learning (Eq. (3.2));

```
// Unsupervised learning
```

for $r = 1$ **to** N_{rd} **do**

  Form neighbourhoods with the current features (Eq. (3.1));

  Curriculum selection of neighbourhoods (Eq. (3.6));

  for iter $= 1$ **to** N_{it} **do**

    Network forward propagation (batch feed-forward);

    Objective loss computation (Eq. (3.7));

    Network back-propagation (Eq. (3.9),(3.10),(3.11));

    Memory feature update (Eq. (3.8));

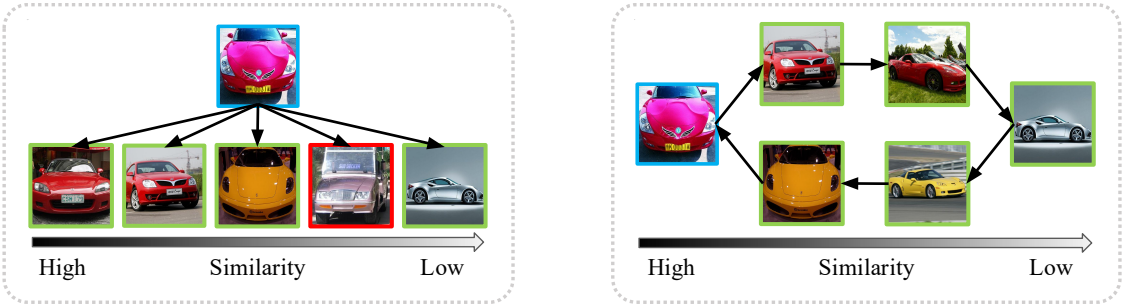
  end

end

3.3 Progressive Affinity Diffusion

In the local data structure formulated in Eq. (3.1), the neighbourhoods represent a kind of data grouping. However, without class label supervision, such structures convey either limited (small k) or unreliable (large k) affinity information between training samples (Figure 3.3 (a)). Using directly $\mathcal{N}_k$ for self-supervision is therefore restricted. To address this problem, we propose affinity diffusion across neighbourhoods. This aims to model the correlation of different $\mathcal{N}_k$ in order to break through their barriers and spread the class identity of one sample through. Conceptually, this leverages global data manifold (Belkin et al. [8]) formed collectively by all neighbourhoods. As each neighbourhood may encode visual affinity information from a distinct perspective, linking them is equivalent to joining different types of variations of a concept, therefore enabling model learning to capture more comprehensive concept boundaries.

An overview of our PAD method is depicted in Figure 3.4. Specifically, PAD is an iterative unsupervised model learning process including three components: **(1)** Affinity graph construction for representing the global structure of training data, **(2)** Affinity diffusion across neighbour-



(a) k-nearest neighbours (b) Strongly connected subgraph

Figure 3.3: An illustration of the effects of neighbourhood's size. **(a)** k-nearest neighbours *vs*. **(b)** strongly connected subgraph. *Blue box*: the anchor. *Green box*: with the same class as the anchor. *Red box*: with a different class against the anchor.

hoods for self-discovering groups of samples with the same semantics, **(3)** Progressive model update by formulating group structure aware objective loss function. They are integrated into a *multi-stage* procedure. In each stage the model mines only the reliable data groups that have emerged thus far in the affinity graph (*i.e.*, *model-matureness-adaptive*) other than clustering all the samples, which then feed into the subsequent model training stages sequentially. We describe the model training details at the end of this section.

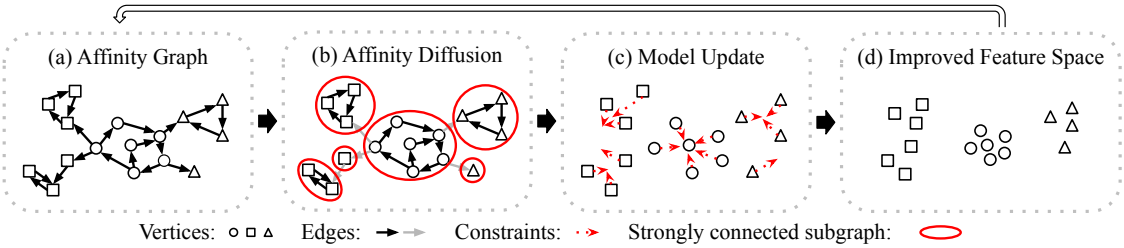


Figure 3.4: An overview of the proposed *Progressive Affinity Diffusion* method. **(a)** Affinity graph construction by k-nearest neighbourhoods. **(b)** Affinity diffusion across neighbourhoods to discover label consistent *strongly connected subgraphs*. **(c)** Progressive model update with self-discovered subgraphs, leading to **(d)** improved feature representations. The model is trained *iteratively*. in a multi-stage procedure.

3.3.1 Affinity diffusion across neighbourhoods

Progressive affinity diffusion is carried out through graphs $\mathcal{G}$ of unlabelled training samples, which is built upon the adjacent relationships among samples encoded in local neighbourhoods (Eq. (3.1)) in the latent feature space. In the directed graphs $\mathcal{G}$, a vertex stands for a sample $\boldsymbol{x}_i$ in the training data X while an edge pointing from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$ indicates that $\boldsymbol{x}_j$ is one of the top-k most similar samples to $\boldsymbol{x}_i$, *i.e.*, $\boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i)$. For discriminative diffusion, it is critical that error introduction and propagation are minimised. To this end, motivated by graph theory (Duda et al.

[44]; Raghavan and Yu [145]), we propose to search **Strongly Connected Subgraphs** (SCS) of the affinity graph $\mathcal{G}$ for revealing underlying semantic concepts. A SCS structure is defined as a set of vertices (images) where each vertex can be reached from any others through neighbouring edges. This means that all vertices of a SCS are highly similar w.r.t. some variation criteria.

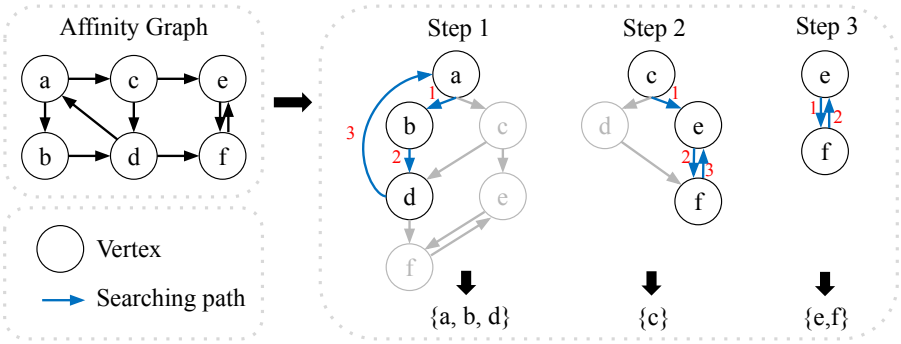


Figure 3.5: An illustration of searching *strongly connected subgraphs* in a neighbourhood affinity graph. In this example, the scale threshold ρ is set to 3.

SCS structures are formed as shown in Figure 3.5. Starting with a random vertex $\boldsymbol{x}_s$, we construct a tree-like structure based on the edges. Then, we traverse the vertices along edges in a depth-first search strategy. We assign nearest neighbours with the highest priority for maximising the average pairwise affinity of a SCS. Depth-first search can easily achieve this condition. It is constrained that each vertex can be visited only one time to avoid repeated traversing. To ensure reachability of any two vertices, we enforce a **cyclic constraint** (Figure 3.3 (b)). In the tree-like structure, one cycle that loops back to the start vertex $\boldsymbol{x}_s$ is necessary, whilst any other cycles allow excluding $\boldsymbol{x}_s$ as long as partly overlapped with the former. In the search process, we aim to find disjoint SCS structures. Hence, we remove all the vertices of a SCS once found from $\mathcal{G}$ to simplify and accelerate the subsequent searching. This process repeats until no SCS exists. Note that at the end, a number of isolated vertices are likely left (outside any SCS).

With the above search method, however, we find that the resulting SCS tends to be over large even when the graph is not dense. This ends up with mixed samples of different class concepts. To address this issue, we further impose an **operational size (scale) threshold** ρ on SCS. This simple constraint works in our case considering that we are *not* seeking a *complete* group of samples per class which can be extremely challenging and risky as in clustering methods, due to complex observation variations. Even splitting a single class into multiple SCS structures, we are still able to obtain a large amount of intra-class variation information, whereas the risk of class mixture is significantly reduced.

3.3.2 Progressive model updating

Next, we describe how the self-discovered SCS structures can be used for progressive model update. For formulation ease, we treat individual samples as special SCS structures each with one sample. We propose to further convert affinity measurements into probability distributions, so that maximum likelihood-based learning objective functions (Goodfellow et al. [58]) can be adopted. Similar as in Section 3.2, we define the probability that any two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are drawn from the same class according to the normalised distance between their features over a stochastic subset of the training data (Eq. (3.2)). For clarity concern, we reiterate the formulation here:

$$p_{i,j} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{x}_j/\tau)}{\sum_{k=1}^{N} \exp(\boldsymbol{x}_i^\top \boldsymbol{x}_k/\tau)} \tag{3.12}$$

where τ is a temperature parameter controlling the distribution concentration degree (Hinton et al. [74]). This quantity is naturally compatible with SCS formation, both relying on pairwise affinity. To reinforce the SCS structural information into model learning, we maximise the same-class possibility of samples per SCS. We therefore formulate a group (subgraph) structure aware objective function as:

$$\mathcal{L}_{\text{scs}} = -\frac{1}{n_{bs}} \sum_{i=1}^{n_{bs}} \log\Big(\sum_{j \in \mathcal{C}(\boldsymbol{x}_i)} p_{i,j} \Big) \tag{3.13}$$

where n_{bs} specifies the mini-batch size and $\mathcal{C}(\boldsymbol{x}_i)$ the samples of the SCS structure including $\boldsymbol{x}_i$. This encourages affinity maximisation of samples within SCS. For isolated samples, $\mathcal{C}(\boldsymbol{x}_i) = \{\boldsymbol{x}_i\}$, $\mathcal{L}_{\text{scs}}$ turns to the instance loss function. The objective loss function formulated in Eq. 3.13 is consistent with that in Eq. 3.7 but the group structures are mined adaptively along the training process of PAD rather than defined manually as independent curriculums in AND. Eq. (3.13) is also analogous to Neighbourhood Component Analysis (Goldberger et al. [56]) when $\mathcal{C}(\boldsymbol{x}_i)$ is replaced with labelled sets.

With summation $\sum_{j \in \mathcal{C}(\boldsymbol{x}_i)} p_{i,j}$, one possible weakness of Eq. (3.13) is that less similar neighbours can be overwhelmed due to over small quantity. To calibrate their importance, we further introduce a **Hard Positive Enhancement** (HPE) strategy. Given a SCS, for a sample $\boldsymbol{x_i}$ we define its hard positive sample as the one $\boldsymbol{x}_{i'}$ with the smallest affinity. In the case of isolated sample when $\mathcal{N}_k(\boldsymbol{x}_i) = \{\boldsymbol{x}_i\}$, we use a randomly transformed variant as its hard positive. For calibration, we minimise the Kullback-Leibler (KL) divergence of model predictions of $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$ as:

$$\mathcal{L}_{\text{hpe}} = \frac{1}{n_{bs}} \sum_{i=1}^{n_{bs}} \sum_{j=1}^{N} p_{i,j} \log \frac{p_{i,j}}{p_{i',j}} \tag{3.14}$$

We obtain the final loss function of the PAD model by weighted summation as:

$$\mathcal{L}_{\text{pad}} = \mathcal{L}_{\text{scs}} + \lambda \mathcal{L}_{\text{hpe}} \tag{3.15}$$

where the weight λ modulates their importance balance. The learnable parameters θ in the model will then be optimised by minimising $\mathcal{L}_{\text{pad}}$.

3.3.3 Model Training

Different from the AND model, thanks to the model-matureness-adaptive design, PAD starts with a randomly initialised CNN rather than pretraining with the sample specificity learning objective (Eq. (3.2)). For efficiency, we update the affinity graph and SCS per epoch, and use a memory to keep track of per-sample representations required by loss computation (Eq. (3.15)). The memory is updated per batch by exponential moving average (Lucas and Saccucci [114]) as formulated in Eq. (3.8). The whole training procedure is summarised in Algorithm 2.

Algorithm 2: PAD for discriminative representation learning.

Input: Training data $\mathcal{I}$, training epochs N_{ep}, iterations per epoch N_{it}.

Output: A class discriminative CNN feature representation model.

for epoch $= 1$ **to** N_{ep} **do**

  Construct the kNN based affinity graph (Eq. (3.1));

  Search strongly connected subgraphs (Figure 3.5);

  for iter $= 1$ **to** N_{it} **do**

    Mini-batch feed-forward through the network;

    Objective loss computation (Eq. (3.15));

    Network back-propagation and model weights update;

    Memory feature refreshing (Eq. (3.8));

  end

end

3.4 Experiments and Evaluations

3.4.1 Datasets, Protocols and Metrics

Datasets. We used 5 image classification benchmarks for evaluating our models. Examples are shown in Figure 3.6. **CIFAR-10/CIFAR-100** (Krizhevsky and Hinton [92]): An image dataset

with 50,000/10,000 train/test images from 10 (/100) object classes. Each class has 6,000 (/600) images with size 32×32. **SVHN** (Netzer et al. [125]): A Street View House Numbers dataset including 10 classes of digit images. **ImageNet** (Russakovsky et al. [154]): A large 1,000 classes object dataset with 1.2 million images for training and 50,000 for test. **STL-10**: An ImageNet adapted dataset containing 500/800 train/test samples from 10 classes as well as 100,000 unlabelled images from auxiliary unknown classes. **MNIST**: A hand-written digits dataset with 60,000/10,000 train/test images from 10 digit classes.

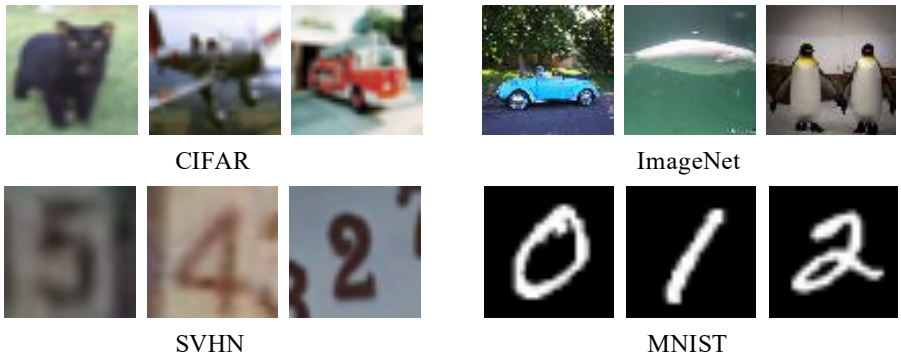


Figure 3.6: Examples of datasets used in neighbourhood mining.

For learning any unsupervised representation model, we assumed and used only the image data but *no* class labels. Unless stated otherwise, we adopted the AlexNet (Krizhevsky et al. [93]) as the neural network architecture for fair comparisons with the state-of-the-art methods. To assess the quality of a learned model for classification at test time, we utilised the ground-truth class labels of the training images *merely* for enabling image categorisation. This does not change the feature representations derived in unsupervised learning.

Protocols. We considered two test protocols for unsupervised learning of discriminative representations. **(1)** *Image classification* (Caron et al. [15]; Wu et al. [195]) where the ground-truth class labels of the training images are used to learn additional classifiers based on *frozen* feature representations. Note that labels are used only for testing. We tested two classification models including Linear Classifier (LC) on top of the last convolutional layer in our backbone and Weighted kNN based on the last fully-connected (FC) features. LC was realised by a fully connected layer. The kNN classifier is based on weighted voting of top-k neighbours $\mathcal{N}_k$ of sample $\boldsymbol{x}_i$ as $p_{i,j} = \sum_{j' \in \mathcal{N}_k} \mathbb{1}[j = y_{j'}] \cdot w_{j'}$ where $p_{i,j}$ is the predicted probability of $\boldsymbol{x}_i$ being drawn from the j-th class while $\mathbb{1}[j = y_{j'}]$ is an identity function returning 1 if label $j = y_{j'}$, and 0 otherwise. We computed the weight $w_{j'}$ as $w_{j'} = \exp(\cos(\boldsymbol{x}_i, \boldsymbol{x}_{j'})/\tau)$ with $\tau = 0.1$ the temperature parameter

and $k = 200$ as in (Wu et al. [195]). Without an additional parametric classifier post-learning involved, kNN reflects *directly* the discriminative capability of the learned feature representations within local areas in the latent space. **(2)** *Image clustering* (Ji et al. [85]; Xie et al. [200]) where k-means is applied if cluster assignments were not explicitly learned by the models for clustering the samples represented by any unsupervised model into the ground-truth number of clusters. Especially, both the training and testing data are used for model learning in the standard clustering setting, unlike the classification setting where only the training data is used.

Evaluation metrics. To measure model's performance, for image classification, we follow the common recipe to test the classifier by the Top-1 Accuracy, namely the proportion of test images that are correctly classified by the classifier learned with the manual labels on the training set. For image clustering, we followed (Ji et al. [85]) to adopt the Clustering Accuracy (Xie et al. [200]) to measure the proportion of samples correctly grouped. Specifically, the clusters yielded by k-means upon the learned feature spaces are associated with the manually defined categories using the Hungarian algorithm (Kuhn [94]).

3.4.2 Implementation Details

For fair comparisons with existing works, we used the same experimental setting as (Wu et al. [195]; Bojanowski and Joulin [13]). We adopted Stochastic Gradient descent (SGD) with Nesterov momentum at 0.9 as the optimiser to train the proposed models for 200 epochs. The learning rate was initiated to be 0.03 and scaled down by a factor of 0.1 every 40 epochs after the first 80. The length of the learned features (outputs of the last FC layer) was set to be 128 and the batch size was 256 for ImageNet and 128 for other datasets. We adopted several common data augmentation techniques same as in (Wang and Gupta [188]) including random horizontal flipping, cropping and colour jittering. We set $\eta = 0.5$ in Eq. (3.8) to update the features memory. For the AND model, we empirically set the number of curriculums to be 4 and the size of the anchor neighbourhoods was set to be $k = 1$ (Eq. (3.1)) for exploring the most local neighbourhoods. For the PAD model, we set $k = 5$ in Eq. (3.1) for affinity graph construction. The maximum size ρ of SCS was set to be 10 and the weight of the learning objective for hard sample mining λ in Eq. (3.15) was 0.8. In practice, the SCS searching is conducted on memory features for efficiency concern and the memory bank takes around 600MB for ImageNet (the largest datasets discussed in this thesis with 1.2M images). Implemented in Tarjan framework (Tarjan [176]),

the worst-case time complexity of our SCS searching is $O(N^2)$ where N is the number of samples. In practice, it takes 4 minutes per epoch to compute the SCS on ImageNet with the overall training time in 4 Tesla P100 GPUs being 6 days. The constructions and updates of SCS is the only computation overhead to the baseline instance discrimination model (Wu et al. [195]). It should not be an efficiency bottleneck. All our parameters were tuned on CIFAR-10 and applied to all the other datasets due to no labelled validation set for cross-validation in unsupervised learning. By using a single setting for all the experiments, we tested our models' scalability and generalisation.

Dataset	CIFAR-10	CIFAR-100	SVHN	ImageNet
Classifier/Feature	Weighted *k*NN / FC			
Random	34.5	12.1	56.8	3.5
Supervised	91.9	69.7	96.5	-
DeepCluster	62.3	22.7	84.9	26.8
Instance	60.3	32.7	79.8	31.3
RotNet	72.5	32.1	77.5	9.2
AND	74.8	41.5	90.9	31.3
PAD	**81.5**	**48.7**	**91.2**	**35.1**
Classifier/Feature	Linear Classifier / conv5			
Random	67.3	32.7	79.2	14.1
Supervised	91.8	71.0	96.1	14.1
DeepCluster	77.9	41.9	92.0	38.2
Instance	70.1	39.4	89.3	35.6
RotNet	84.1	57.4	92.3	36.5
AND	77.6	47.9	**93.7**	37.8
PAD	**84.7**	**58.6**	93.2	**38.6**

Table 3.1: Comparisons with the state-of-the-art methods in unsupervised image classification. The 1st/2nd best results were marked in **red**/blue. Results of previous methods were reproduced by us using the implementations released by their authors.

3.4.3 Comparisons with the State-of-the-Art

Image classification. In Table 3.1, we compared our AND and PAD methods with the representative works of *clustering analysis* (DeepCluster (Caron et al. [15]), *self-supervised learning* (RotNet (Gidaris et al. [55])), and *instance contrastive learning* (Instance (Wu et al. [195])) as well as the random and supervised learning baselines on four benchmarks. We make three observations:

- The PAD model surpasses all competitors under either classification model, often by a large margin while the AND model is superior when testing with the kNN classifier. This suggests the performance advantages of our methods thanks to its strong capability of discovering semantic consistent neighbourhood structures.

- When compared with kNN classifier, linear classifier tends to yield better results due to using extra parameters. This is particularly so for the pretext task based model RotNet. This is because the pretext task is less relevant to classification, leading to weaker representation as compared to grouping based methods like AND and PAD. In contrast, the margins obtained by our models with kNN as the classifier tend to be larger than those by linear classifier. This indicates features derived by our models are favourably more ready for direct use without extra classifier training as post-processing, *i.e*. more discriminative regarding the semantic classes.

- As an intermediate representation between clusters and instances, tiny neighbourhoods are exploited in AND for revealing class boundaries and improves the results in most cases. However, this method is restricted by the small size of neighbourhoods. PAD addresses this limitation by affinity diffusion across adjacent neighbourhoods.

Image clustering. Apart from sample-wise image classification, we further validated our models in image clustering which reflects the representation quality in describing global data structures. We compared PAD with two groups of alternative methods, **(1)** *Clustering methods*: JULE (Yang et al. [207]), DEC (Xie et al. [200]), DAC (Chang et al. [20]), ADC (Haeusser et al. [63]) and IIC (Ji et al. [85]); **(2)** *Generic representation learning methods*: Triplets (Schultz and Joachims [157]), AE (Bengio et al. [10]), Sparse AE (Ng [126]), Denoising AE (Vincent et al. [184]), Variational Bayes AE (Kingma and Welling [90]), SWWAE (Zhao et al. [225]),

Methods	MNIST	STL-10	CIFAR-10	CIFAR-100
JULE	96.4	27.7	27.2	13.7
DEC	84.3	35.9	30.1	18.5
DAC	97.8	47.0	52.2	23.8
ADC	99.2	52.0	32.5	16.0
IIC	98.4	59.8	57.6	25.5
Random†	48.1	20.1	18.6	10.3
Triplets†	52.5	24.4	20.5	9.9
AE†	81.2	30.3	31.4	16.5
Sparse AE†	82.7	32.0	29.7	15.7
Denosing AE†	83.2	30.2	29.7	15.1
Variational Bayes AE†	83.2	28.2	29.1	15.2
SWWAE†	82.5	27.0	28.4	14.7
DCGAN†	82.8	29.8	31.5	15.1
DeepCluster†	65.6	33.4	37.4	18.9
PAD†	**98.2**	**46.5**	**62.6**	**28.8**

Table 3.2: Comparisons with the state-of-the-art methods in unsupervised image clustering. Methods with † generate cluster assignments with the help of k-means. The 1st/2nd best results were marked in **red**/blue. Results of previous methods were adopted from (Ji et al. [85]).

DCGAN (Radford et al. [143]) and DeepCluster (Caron et al. [15]). For the latter group including PAD, we further applied k-means to generate their clustering solutions. For PAD, we reported the average result over 10 runs. For competitors, we used the results from (Ji et al. [85]). Despite different modelling purposes, we performed both within and cross group comparisons. We made a couple of observations from Table 3.2:

- The first group of methods tends to produce higher clustering results, thanks to their joint learning of feature representations and clustering by using the ground-truth class number prior in end-to-end model training, *i.e.*, consistent between training and test objectives. Among them, IIC (Ji et al. [85]) achieves the best results while JULE (Yang et al. [207]) serving as a baseline that share a similar agglomerative clustering idea with PAD. .

- Without taking clustering as objective, the second group of methods is relatively inferior in modelling data group structures. However, PAD again reaches the best performance consistently in this group. Crucially, our model is on par with all the dedicated clustering methods and even surpasses them on CIFAR-10 and CIFAR-100 with significant margins, regardless of the disadvantage on STL-10 which, we conjugate, is due to some distracting impact from auxiliary unknown categories. This indicates the efficacy of our unsupervised learning method in capturing the holistic data distribution. We attribute this advantage to the favourable ability of our method in seeking the latent class consistent groups with high variations of individual concepts.

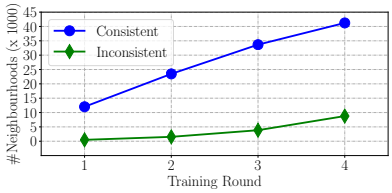


Figure 3.7: Neighbourhood quality of AND over rounds on CIFAR-10.

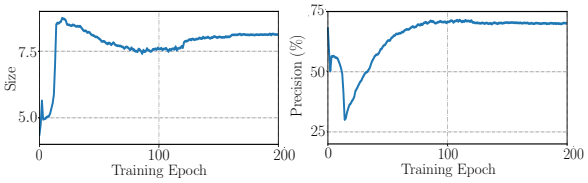


Figure 3.8: Statistics dynamics of SCS during training on CIFAR-10: average size and precision.

Neighbourhood quality. We examined the class consistency of local neighbourhoods discovered throughout the training process. Figure 3.7 shows that the numbers of both class-consistent and inconsistent anchor neighbourhoods increase along with the training rounds, and more importantly the consistent ones raise much more rapidly. This explains the performance advantages of the AND model and the benefit of exploring progressive curriculum learning. Besides, we examined the training dynamics of SCS size and precision in PAD. Figure 3.8 shows that PAD starts with finding small SCS structures due to weak representation power, then explores larger ones at decreasing precision, and finally converges the size whilst increases the precision before both levels off. High precision of SCS is a key for enabling more discriminative unsupervised learning by PAD.

3.4.4 Component Analysis of AND

We conducted detailed component analysis with the weighted kNN classifier and FC features to provide insights into the AND model.

Backbone network. We tested the generalisation of AND with varying-capacity networks. We further evaluated ResNet18 and ResNet101 (He et al. [68]) on CIFAR-10. Table 3.3 shows that

AND benefits from stronger net architectures. A similar observation was made on ImageNet: 41.2% (ResNet18) vs. 31.3% (AlexNet).

Network	AlexNet	ResNet18	ResNet101
Accuracy	74.8	86.3	**88.4**

Table 3.3: Network generalisation analysis of AND on CIFAR-10.

Neighbourhood size. Neighbourhood size is an important parameter in AND since it controls label consistency of neighbourhoods and finally the model performance. We evaluated its effect using ResNet18 on CIFAR-10 by varying k from 1 (the default value) to 100. Note that $k = 1$ refers to the model trained with neighbourhoods with at most *two* member samples. Figure 3.9 shows that the smallest neighbourhoods (*i.e.*, $k = 1$) are the best choice. This implies high variety of imagery data, so smaller neighbourhoods are preferred if without adaptive and reliable strategy for neighbourhoods expansion.

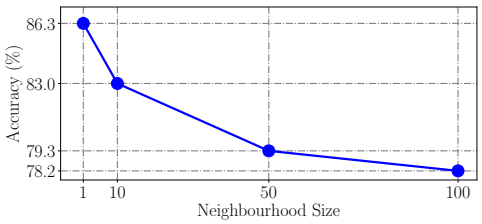


Figure 3.9: Effect of the neighbourhood size in AND on CIFAR-10.

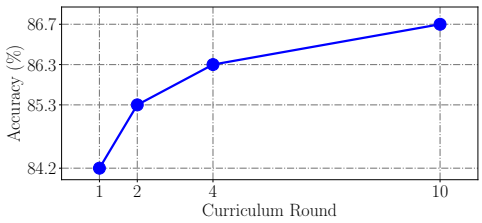


Figure 3.10: Effect of the curriculum round in AND on CIFAR-10.

Curriculum round. We tested the effect of the curriculum round (R in Eq. (3.6)) of progressive neighbourhood discovery. More rounds consume higher training costs. Figure 3.10 shows that using 4 rounds gives a good trade-off between model training efficiency and feature performance. Often, per-round epoch number N_{ep} affects the training efficiency and performance. We investigated its effect and found that AND achieves 83.3% by N_{ep}=50 vs. 84.8% by N_{ep}=100.

One-off *vs*. curriculum neighbourhood discovery. We evaluated the benefit of AND's curriculum. To this end, we compared with the *one-off* discovery counterpart where all anchor neighbourhoods are exploited one time. Table 3.4 shows that the proposed multi-round progressive discovery via a curriculum is effective to discover more reliable anchor neighbourhoods for superior unsupervised learning.

Neighbourhood Discovery	One-off Exploration	Curriculum Learning
Accuracy	84.2	**86.3**

Table 3.4: One-off vs. curriculum discovery in AND on CIFAR-10.

3.4.5 Component Analysis of PAD

Similarly, we conducted a sequence of detailed component evaluations and performance analysis for the PAD model on the image classification task with kNN as classifier.

Effect of affinity diffusion. To investigate the effect of affinity diffusion, we tested the performance of models in which $\mathcal{C}(\boldsymbol{x})$ for each sample was replaced by its k-nearest neighbours $\mathcal{N}_k(\boldsymbol{x})$ as shown in Figure 3.3 (a) and the size k was set to 10. According to Table 3.5, discovering consistent sample groups according to affinity diffusion across adjacent neighbourhoods with necessary constraints clearly benefits the discriminative learning of models in all cases.

Diffusion	CIFAR-10	CIFAR-100	SVHN
✗	77.5	34.5	89.5
✓	**81.5**	**48.7**	**91.2**

Table 3.5: Effect of affinity diffusion in PAD.

Cyclic and scale constraints. We examined the effect of *cyclic* and *scale* constraints (Section 3.3.1) and observed from Table 3.6 that: **(1)** Cyclic constraint brings consistently performance gain, particularly in the most challenging CIFAR-100 test. This is because of the presence of subtle visual discrepancy between fine-grained classes, which leads to more wrong association in affinity diffusion. **(2)** Scale constraint is clearly necessary for ensuring the effectiveness of our model in all cases. Without it, different classes would be mixed up in diffusion due to complex visual patterns exhibited in images.

Hard positive enhancement. Table 3.7 (left) shows that constraining the prediction between hard positive pairs is clearly beneficial for model discriminative learning. This confirms the overwhelming effect among within-SCS samples when using normalised affinity measurements (Eq. (3.2)) to quantify loss function (Eq. (3.13)), and suggests the efficacy of our enhancement

Cyclic	CIFAR-10	CIFAR-100	SVHN
✗	73.3	30.6	91.1
✓	81.5	48.7	91.2

Scale	CIFAR-10	CIFAR-100	SVHN
✗	20.2	1.8	20.3
✓	81.5	48.7	91.2

Table 3.6: Effect of **(left)** cyclic and **(right)** scale constraints in forming SCS structures.

strategy. We also compared two loss designs: Feature Cosine Similarity (FCS) *vs*. KL divergence. Table 3.7 (right) suggests the superiority of KL over FCS. A plausible explanation is that KL can integrate with SCS loss $\mathcal{L}_{\text{scs}}$ (Eq. (3.13)) in a more harmonious manner, as both are based on class posterior probability measurements.

HPE	CIFAR-10	CIFAR-100	SVHN
✗	69.8	30.9	80.9
✓	81.5	48.7	91.2

Design	CIFAR-10	CIFAR-100	SVHN
FCS	72.7	39.1	90.4
KL	81.5	48.7	91.2

Table 3.7: (**Left**) Effect of hard positive enhancement and (**Right**) the HPE loss design comparison of Feature Cosine Similarity and Kullback-Leibler divergence.

Parameter analysis. We evaluated 3 parameters of PAD on CIFAR-10: **(1)** Affinity graph density k, **(2)** SCS scale threshold ρ, and **(3)** weight λ of hard positive enhancement loss. Table 3.8 shows that the parameters are insensitive with a wide range of good values, indicating training robustness.

Parameter	Affinity graph density k				SCS scale threshold ρ				HPE loss weight λ			
Value	1	3	5	10	5	10	50	100	0.2	0.5	0.8	1.0
Accuracy	77.8	79.3	80.4	79.5	79.3	80.4	80.3	79.7	78.2	79.9	80.4	80.2

Table 3.8: Model parameter analysis of PAD on CIFAR-10. **Left**: Affinity graph sparsity k; **Middle**: SCS scale threshold ρ; **Right**: Weight λ of hard positive enhancement loss.

Computation cost of SCS searching. To validate the complexity of SCS searching, we tested the searching time on ImageNet: 4 mins per epoch, 800 mins among the overall 6 days training.

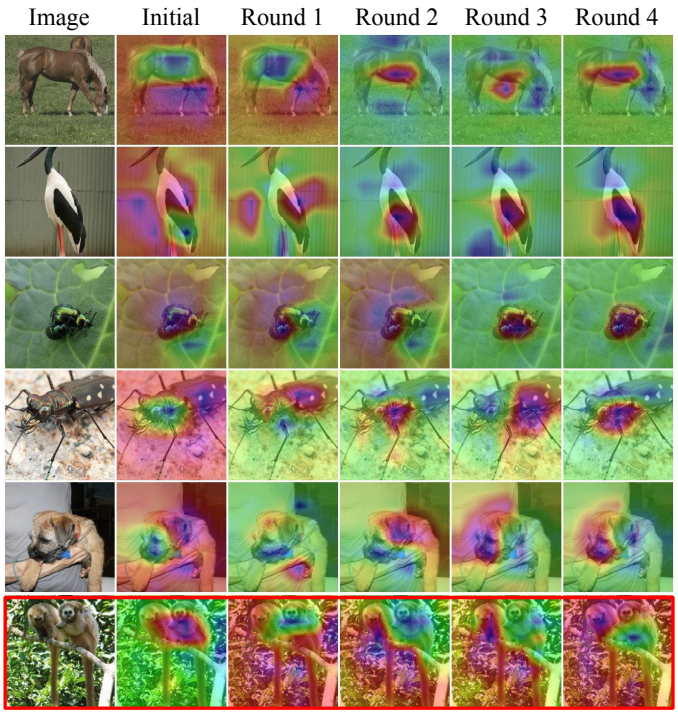


Figure 3.11: The evolving dynamics of the AND model's learning attention throughout the training rounds on six ImageNet classes. Red bounding box indicates a failure case.

3.4.6 Visualisation and Qualitative Study

To further understand how the idea of local neighbourhood discovery benefits feature representation learning, we tracked the AND's modelling attention by Grad-Cam (Selvaraju et al. [158]) to visualise which parts of training images the model is focusing on throughout the curriculum rounds. We have the following observations from Figure 3.11: **(1)** Often the model initially looks at class irrelevant image regions. This suggests that sample specificity is a less effective supervision signal for guiding model class-discriminative training. **(2)** In cases, the AND model is able to gradually shift the learning attention towards the class relevant parts therefore yielding a more discriminative model. **(3)** The AND may fail to capture the object attention, *e.g.* due to cluttered background and poor lighting condition.

What's more, to provide visual interpretation of how the PAD model yielded more discriminative representations than that of AND by benefiting from discovering neighbourhoods in larger sizes, we conducted a case study of affinity diffusion on STL-10. Figure 3.12 shows that when the model is immature (at the early training stage), wrong cross-class diffusion may happen frequently; the cyclic constraint can help detect this and early stop error accumulation. Besides, it is shown that SCS can better capture semantic similarity beyond pairwise affinity measurements (see dashed curve). However, as expected not all the hard positive pairs are discovered due to

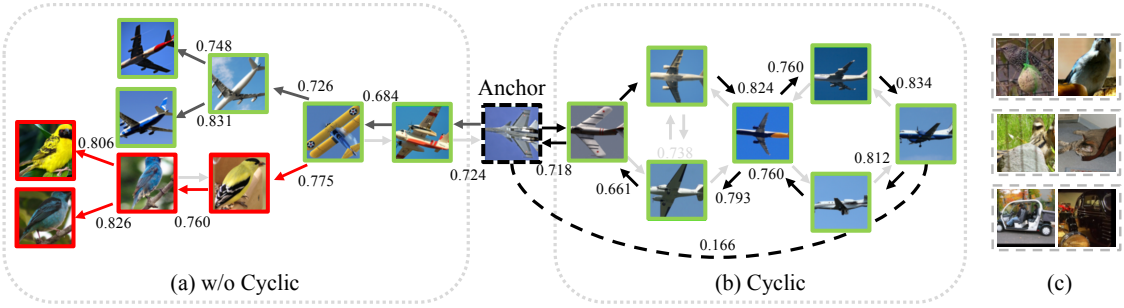


Figure 3.12: Case studies of PAD on STL-10. **(a)** *vs*. **(b)**: Cyclic constraint helps detect erroneous affinity diffusion. **(c)** Extremely hard positive pairs undiscovered by PAD. *Red solid box*: with a different class as the anchor; *Green solid box*: with the same class as the anchor. The numbers above arrows are the corresponding pairwise affinity scores.

extreme viewing condition discrepancy as shown in Figure 3.12 (c).

3.5 Summary

In this chapter, we presented two novel approaches (Anchor Neighbourhood Discovery and Progressive Affinity Diffusion) for unsupervised learning of class discriminative feature representations for images through class consistent neighbourhood discovery and supervision in a progressive manner. With the AND model, we make a first attempt to avoid the notorious grouping noises whilst still preserving the intrinsic merits of clustering for effective inference of the latent class decision boundaries. It is also superior to the existing sample-specific instance contrastive learning strategy, due to the unique capability of propagating the self-discovered sample-to-sample class relationship information in end-to-end model optimisation. Moreover, to further expand the size of class-consistent local neighbourhoods in order to introduce more complex intra-class image variations to encourage the visual features to be more class-sensitive, we further present PAD with a model-matureness-adaptive design. It is achieved by self-discovering class consistent strongly connected subgraphs in neighbourhood affinity graphs and formulating group structure aware objective loss function. Critically, The PAD model overcomes the small locality limitation of neighbourhoods in AND, whilst preserving and integrating its intrinsic strengths for more effective discriminative learning. Extensive experiments on five visual recognition benchmark datasets on image classification and image clustering tasks validate the superiority of both AND and PAD over a wide spectrum of state-of-the-art unsupervised deep learning methods. In-depth component analysis and intuitive qualitative cases study were provided to give insights on the model advantages of both the formulations.

Chapter 4

Discovering Knowledge from Global Data Structure in Deep Clustering

---

This chapter investigates knowledge mining from global data structure to learn jointly discriminative feature representation of visual data as well as their semantic partitions without accessing any manual class labels of target data. The image clusters are considered semantically plausible if they can be mapped to the human defined categories *one-to-one*. In this case, images of the same clusters can be described by the labels of the matched classes, which are usually human-understandable short phrases like cat, dog, ship, etc. The task studied in this Chapter is different from the one in Chapter 3 by predicting the probabilities of samples being from a set of clusters as model's outputs while the latter producing feature representations instead. Specifically, given a set of N *unlabelled* target images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_N\}$ drawn from C semantic classes $\mathcal{Y} = \{y_1, y_2, \cdots, y_N \mid \forall i \in [1,N],\ y_i \in [1,C]\}$, the objective of deep clustering is learning to separate $\mathcal{I}$ into C clusters in an unsupervised manner by a CNN model. There are typically two components learned jointly end-to-end: **(1)** a feature extractor $f_{\boldsymbol{\theta}}(\cdot)$ that maps the target images into vector representations: $\boldsymbol{x}_i = f_{\boldsymbol{\theta}}(\boldsymbol{I}_i)$, and **(2)** a classifier $g_{\boldsymbol{\phi}}(\cdot)$ that assigns each feature representation $\boldsymbol{x}_i$ with a cluster membership distribution: $\boldsymbol{p}_i = g_{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \{p_{i,1}; p_{i,2}; \cdots; p_{i,C}\}$. Once the model is trained, the cluster assignment can be predicted in a maximum likelihood manner as:

$$\tilde{y}_i = \arg\max_c (p_{i,c}),\ \ c \in \{1,2,\cdots,C\}. \tag{4.1}$$

Unlike in Chapter 3 where $p_{i,j}$ denotes the probability that the i-th and j-th samples are positive, $p_{i,j}$ indicates how likely the i-th sample is from the j-th cluster here. Ideally, all the samples

of a cluster would share the same target class labels. That being said, we aim to discover the underlying semantic class decision boundary directly from the raw data samples.

Aiming at the two challenges of joint learning data partitions and feature representations summarised in Section 1.3, this chapter is organised as follows: Section 4.1 reiterates the challenge of learning semantically plausible data partitions without the constraints from human annotations to motivate the novel PartItion Confidence mAximisation model presented in this chapter with its overview provided in Section 4.1.1 and detailed formulations given in Section 4.1.2 and model training in Section 4.1.3; Section 4.2 discusses the obstacles of error-propagation when benefiting deep clustering by the powerful unsupervised learning techniques for generic visual representations, which motivates a new variant of instance contrastive learning (He et al. [69]) to derive class-discriminative feature representations by Semantic Contrastive Learning. An overview of SCL is given in 4.2.1 while the model designs and training strategy are in Section 4.2.2 and Section 4.2.3, respectively; Section 4.3 provides extensive experiments for evaluating the two proposed models on a wide range of natural object recognition benchmark datasets with in-depth analysis and comprehensive ablation studies. Section 4.4 summaries the ideas of discovering knowledge from global data structure studied here.

4.1 Maximising Semantic Plausibility of Clusters

In general, image clustering is *not* a well-defined problem as multiple different solutions can all make sense of the input data (Xu and Wunsch [204]). This makes deep clustering extremely challenging due to totally lacking high-level guidance knowledge from human annotations. However, in object-centric images (*e.g*. ImageNet (Deng et al. [37])), the objects-of-interest that potentially drawing human's attention are usually occupying also the most salient parts of images. Therefore, regardless the existence of intra-class visual discrepancy and inter-class affinity, the majority of samples from the same semantic classes are still expected to share a high proportion of visual information. In this case, although a set of imagery data can be separated in numerous ways according to various criteria, assigning samples of the same semantic categories to different clusters will implicitly reduce the resulted within-cluster compactness and between-cluster separability (Wu et al. [195]), leading to lower partition confidence. Therefore, we assume that the most *confident* data partition is the most promising clustering solution we are seeking for, which is likely to be mapped to the human-defined object categories one-to-one, *i.e*. semantic plausible.

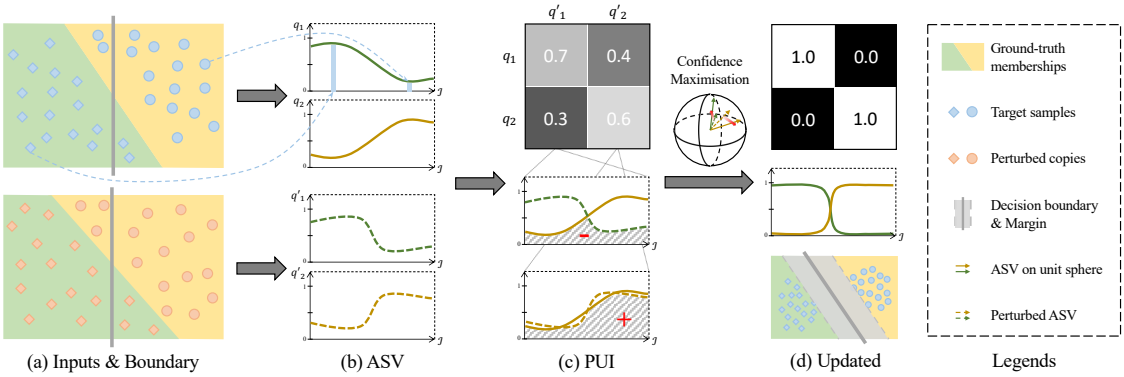


Figure 4.1: An overview of the proposed PartItion Confidence mAximisation method.

4.1.1 Approach Overview

In this chapter, we formulate a novel deep clustering method, called *PartItion Confidence mAximisation*. PICA is based on a Partition Uncertainty Index (PUI) that measures how a deep CNN is capable of interpreting and partitioning the target image data. This index is differentiable, hence, PICA simply needs to optimise it without whistles and bells using any off-the-shelf CNN models. An overview of PICA is depicted in Figure 4.1. Given the input data and the decision boundaries determined by the CNN model (Figure 4.1 (a)), the PICA model computes the *cluster-wise Assignment Statistics Vector* (ASV) in the forward pass using a mini-batch data (Figure 4.1 (b)). To minimise the *partition uncertainty index* depicted in Figure 4.1 (c), the PICA model is trained to discriminate the ASV of all negative cluster pairs on the hypersphere through a dedicated objective loss function (Figure 4.1 (d)), so as to learn the most confident and potentially promising clustering solution. Moreover, to ensure model's invariance to visual transformations, the PUI is computed by a batch of samples and its perturbed copy obtained by random data augmentations. Such a design leads to the dual-branch architecture of PICA.

4.1.2 Partition Uncertainty Index

We start by formulating a partition uncertainty index, a key element of our PICA. Given an input image $\boldsymbol{I}_i$, suppose the cluster prediction of a CNN model is denoted as:

$$\boldsymbol{p}_i = \begin{bmatrix} \boldsymbol{p}_{i,1} \\ \cdots \\ \boldsymbol{p}_{i,C} \end{bmatrix} \in \mathbb{R}^{C\times 1}, \tag{4.2}$$

where $p_{i,j}$ specifies the predicted probability of the i-th image being assigned to the j-th cluster, and there are a total of C clusters ($j \in [1,2,\cdots,C]$). We then obtain the cluster prediction matrix

of all the N target images as

$$P = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N] \in \mathbb{R}^{C \times N}. \tag{4.3}$$

For presentation ease, we denote the j-th row of P as:

$$\boldsymbol{q}_j = [p_{1,j}, p_{2,j}, \cdots, p_{N,j}] \in \mathbb{R}^{1 \times N}, \;\; j \in [1, 2, \cdots, C]. \tag{4.4}$$

Clearly, $\boldsymbol{q}_j$ collects the probability values of all the images for the j-th cluster, which summarises the *assignment statistics* of that cluster over the whole target data. Hence, we call it as a *cluster-wise Assignment Statistics Vector*.

Ideally, each image is assigned to only one cluster, *i.e*. each $\boldsymbol{p}_i$ is a one-hot vector (same as the ground-truth label vectors in supervised image classification). It is intuitive that this corresponds to the *most confident* clustering solution which is the objective that PICA aims to achieve. For enabling the learning process of a deep clustering model towards this ideal (most confident) case, an objective loss function is typically needed. To that end, we design a partition uncertainty index as the learning target. Specifically, we observe that in the ideal case above, the ASV quantities of any two clusters, $\boldsymbol{q}_{j_1}$ and $\boldsymbol{q}_{j_2}$, are *orthogonal* to each other. Mathematically, this means that their cosine similarity (Eq. (4.5)) is 0 (due to not negative values in $\boldsymbol{q}_{j_1}$ and $\boldsymbol{q}_{j_2}$).

$$\cos(\boldsymbol{q}_{j_1}, \boldsymbol{q}_{j_2}) = \frac{\boldsymbol{q}_{j_1} \cdot \boldsymbol{q}_{j_2}}{\|\boldsymbol{q}_{j_1}\|_2 \|\boldsymbol{q}_{j_2}\|_2}, \;\; j_1, j_2 \in [1, \cdots, K] \tag{4.5}$$

In the worst clustering cases where all the cluster prediction $\boldsymbol{p}_i$ are the same (*e.g*. the uniform distribution vector), we also have a constant value: $\cos(\boldsymbol{q}_{j_1}, \boldsymbol{q}_{j_2}) = 1$, since all the ASV quantities are the same. For any case in-between, the ASV cosine similarity of two clusters will range from 0 (most confident) to 1 (least confident).

In light of the above analysis, we formulate a partition uncertainty index as the ASV cosine similarity set of all the cluster pairs as:

$$\mathcal{M}_{\text{PUI}}(j_1, j_2) = \cos(\boldsymbol{q}_{j_1}, \boldsymbol{q}_{j_2}), \;\; j_1, j_2 \in [1, C]. \tag{4.6}$$

In form, $\mathcal{M}_{\text{PUI}}$ is a $C \times C$ matrix. By doing so, the learning objective of PICA is then to minimise the PUI (except the diagonal elements), which is supposed to provide the most confident clustering solution at its minimum.

A Stochastic Approximation of PUI. The PUI as formulated in Eq. (4.6) requires using the entire target data which is often at large scale. This renders it *unsuited* to stochastic mini-batch

based deep learning. To address this problem, we propose a stochastic approximation of PUI. Specifically, instead of using all the images (which is deterministic), at each training iteration we use a random subset $\mathcal{I}^t$ of them. In probability theory and statistics, this is sampling from a discrete uniform distribution in the whole target data space (Hogg et al. [78]). We call this approximation as *Stochastic PUI*. In practice, this allows to fit easily the mini-batch training of the standard deep learning, *e.g.* simply setting $\mathcal{I}^t$ as a mini-batch.

Formally, at the t-th training iteration, we have a mini-batch B of n_{bs} samples to train the model and set $\mathcal{I}^t = B$. Let us denote the cluster prediction matrix of $\mathcal{I}^t$ made by the up-to-date model as:

$$\boldsymbol{P}^t = \begin{bmatrix} \boldsymbol{q}_1^t \\ \dots \\ \boldsymbol{q}_C^t \end{bmatrix} \in \mathbb{R}^{C \times n_{\text{bs}}}, \tag{4.7}$$

where $\boldsymbol{q}_j^t \in \mathbb{R}^{1 \times n_{\text{bs}}}$ is the ASV of j-th cluster on mini-batch $\mathcal{I}^t$. As Eq. (4.6), we obtain the Stochastic PUI:

$$\mathcal{M}_{\text{S-PUI}}(j_1, j_2) = \cos(\boldsymbol{q}_{j_1}^t, \boldsymbol{q}_{j_2}^t), \;\; j_1, j_2 \in [1, C]. \tag{4.8}$$

The Stochastic PUI is in a spirit of dropout (Srivastava et al. [166]). Instead of *neurons*, we randomly drop *data samples* in this case and realised in the standard mini-batch sampling process.

4.1.3 Model Training

Learning objective function. Given the Stochastic PUI $\mathcal{M}_{\text{S-PUI}}$, as discussed earlier PICA is then trained to minimise it excluding the diagonal elements. To derive a typical objective loss function, we usually need a *scalar* measure. However, $\mathcal{M}_{\text{S-PUI}}$ is a $C \times C$ matrix. There is hence a need to transform it.

Recall that for any two different clusters, we want to minimise their ASV cosine similarity. This is actually reinforcing self-attention (Vaswani et al. [182]) by treating each cluster as a data sample and suppressing all the inter-sample attention. Hence, we apply a softmax operation as self-attention to each cluster j and obtain a probabilistic measurement as:

$$m_{j,j'} = \frac{\exp(\mathcal{M}_{\text{S-PUI}}(j, j'))}{\sum_{c=1}^{C} \exp(\mathcal{M}_{\text{S-PUI}}(j, c))}, \;\; j' \in [1, C]. \tag{4.9}$$

With this transformation, the learning objective is further simplified into maximising $\{m_{j,j}\}_{j=1}^{C}$.

By further treating $m_{j,j}$ as the model prediction probability on the ground-truth class of a training sample (a cluster j in this context), a natural choice is then to exploit the common cross-entropy loss function:

$$\mathcal{L}_{\text{ce}} = \frac{1}{C}\sum_{j=1}^{C} -\log(m_{j,j}). \tag{4.10}$$

As such, we formulate a scalar objective loss function $\mathcal{L}_{\text{ce}}$ that minimises effectively the matrix $\mathcal{M}_{\text{S-PUI}}$.

In clustering, there are algorithm-agnostic trivial solutions that assign a majority of samples into a minority of clusters. To avoid this, we introduce an extra constraint that minimises *negative* entropy of the cluster size distribution:

$$\mathcal{L}_{\text{ne}} = \log(C) - H(\mathcal{Z}), \text{ with } \mathcal{Z} = [z_1, z_2, \cdots, z_K] \tag{4.11}$$

where $H(\cdot)$ is the entropy of a distribution, and $\mathcal{Z}$ is L_1 normalised soft cluster size distribution with each element computed as $z_j = \frac{\Sigma(\boldsymbol{q}_j^t)}{\sum_{c=1}^{C}\Sigma(\boldsymbol{q}_c^t)}$. With the maximal entropy of a C-dimensional probability distribution, $\log(C)$ is to ensure non-negative loss values.

The overall objective function of PICA is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{ne}} \tag{4.12}$$

where λ is a weight parameter. The objective function (Eq. (4.12)) of PICA is differentiable end-to-end, enabling the conventional stochastic gradient descent algorithm for model training. To improve the model robustness to visual transformations we use data augmentation to randomly perturb the training data distribution. We enforce the clustering invariance against image perturbations at the global solution level. More specifically, we use the original data to compute $\boldsymbol{q}_{j_1}^t$ and the transformed data to compute $\boldsymbol{q}_{j_2}^t$ in Eq. (4.8) at each iteration. The training procedure is summarised in Algorithm 3.

4.2 Alleviating Error-propagations to Features

Apart from the simultaneous representation learning and clustering paradigm studied in the above section, alternative learning is another widely adopted strategy in deep clustering. This section further investigates the alternation strategy and the error-propagation problem commonly seen in it. A novel approach is introduced to alleviate the negative impacts of error-propagation with the help of instance contrastive learning (He et al. [69]; Chen et al. [27]; Tian et al. [177]). Among

Algorithm 3: PICA for deep clustering.

Input: Unlabelled data $\mathcal{I}$, training iterations N_{it} target cluster number C.

Output: A deep clustering model;

for iter $= 1$ **to** N_{it} **do**

Sampling a random mini-batch of images;

Feeding the mini-batch into the deep model;

Computing per-cluster ASV (Eq. (4.4));

Computing the Stochastic PUI matrix (Eq. (4.8));

Computing the objective loss (Eq. (4.12));

Model weights update by back-propagation;

end

the recent developments of unsupervised learning of generic visual representations which are applicable to various downstreamed tasks with fine-tuning, instance contrastive learning has been shown to excel at learning sample-specific discriminative feature representations by *instance discrimination*, with a direct benefit to unsupervised clustering through encoding visual similarity in feature distance. Specifically, instance contrastive learning derives image features by learning with per-sample pseudo classes generated by global linear data augmentations, which differentiates every independent instance from all or a random subset of training data (negative/contrastive set) regardless of their semantic class memberships. However, such an approach does not optimise *concept* (class) discrimination between clusters and is unaware of any potential nonlinear intra-class variations. Lacking considerations for nonlinear class decision boundaries, the resulting features are limited to apparent visual similarity (*e.g.* pixel intensity) rather than latent semantic membership interpretations. On the other hand, considering that the instance discrimination formulation is agnostic to sample's underlying class memberships, such a learning strategy provides a robust manner to deal with unreliable estimations of cluster assignment when imposing false negative pairwise relationships into contrastive sets. In this regard, we present a new variant of contrastive learning in this section for deriving class discriminative visual representations called Semantic Contrastive Learning. The SCL model simultaneously conducts *cross-cluster* instance discrimination and cluster discrimination to learn jointly visual features and cluster decision boundaries. In the *cross-cluster* instance discrimination, rather than explicitly pushing the

potentially positive samples closer in the latent space to facilitate feature invariance to intra-class visual discrepancy, we implicitly encourage this by assigning a common contrastive set to the pseudo positive samples so to alleviate the errors propagated from the unreliable (noisy) cluster assignments. Meanwhile, the cluster decision boundaries are derived to maximise the consistency between sample's distances in cluster-level (semantic) and instance-level (visual).

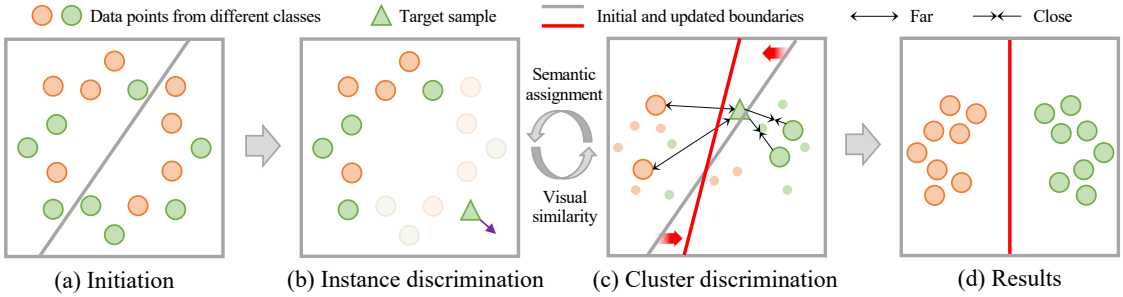


Figure 4.2: An overview of the proposed Semantic Contrastive Learning method.

4.2.1 Approach Overview

An overview of the SCL model is depicted in Figure 4.2. Given a randomly initiated feature space and the decision boundary of a set of clusters (Figure 4.2 (a)), the SCL model learns the visual similarities of instances in each cluster by instance discrimination against samples from other clusters (Figure 4.2 (b)) and optimises each cluster decision boundary by assigning the instances to the cluster with the most similar samples (Figure 4.2 (c)). As shown in Figure 4.2 (d), SCL converges to a consensus between maximising inter-cluster instance-level (visual) diversity and intra-cluster cluster-level (semantic) compactness. This results in semantically more plausible clusters to interpret the underlying semantic concepts without explicitly learning from labels.

4.2.2 Instance and Cluster Discrimination

We start with formulating a new *cross-cluster* instance discrimination learning objective with a novel *semantic memory*. The aim is to learn visual features to be discriminative across clusters and design the semantic memory to facilitate simultaneous instance and cluster discrimination.

Cross-cluster instance discrimination. Our feature learning objective is formulated to differentiate every individual instance against its pseudo-negative samples so to reduce its visual redundancy regarding images of other clusters. Given random partitions at the beginning of training, by isolating samples from different clusters, the model behaves as instance contrastive

learning and outputs per sample-specific visual features. Intuitively, visually similar samples are expected to share more class-specific unique information, their representations will therefore be gradually gathered closer and grouped into the same clusters by our cluster discrimination detailed later. Along the clustering process with increasingly better and stable cluster assignments, the contrastive set of every sample will absorb more visually dissimilar counterparts, instead of random ones. Consequently, the learning objective becomes reducing cross-cluster visual redundancy, resulting in desired features that are aware of inter-cluster visual discrepancies and invariant within clusters.

Our cross-cluster instance discrimination is a generic formulation, integratable with existing instance contrastive learning methods. The main refinement involved is to add the component of contrastive sets at the cluster level in a plug-in manner. We take the MoCo (He et al. [69]; Chen et al. [29]) as an example.

To be concrete, we first formulate a mapping function $f_{\boldsymbol{\theta}}$ from a pixel space to a representational space as an encoder with learnable weights $\boldsymbol{\theta}$. Similarly, we construct another momentum encoder $f_{\hat{\boldsymbol{\theta}}}$ with an identical structure but independent parameters $\hat{\boldsymbol{\theta}}$. Given an unlabelled dataset $\mathcal{I}$, we randomly apply a set of transformations $\mathcal{T}$ to each image for distribution perturbation. We then represent two perturbed copies of each instance, $\mathcal{T}_1(\boldsymbol{I}_i)$ and $\mathcal{T}_2(\boldsymbol{I}_i)$, by the encoder and momentum encoder respectively and denote them as $\boldsymbol{x}_i = f_{\boldsymbol{\theta}}(\mathcal{T}_1(\boldsymbol{I}_i))$ and $\hat{\boldsymbol{x}}_i = f_{\hat{\boldsymbol{\theta}}}(\mathcal{T}_2(\boldsymbol{I}_i))$. Given the pseudo labels of all the samples $\mathcal{Y} = \{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_N\},\ \tilde{y}_i \in [1, C]$ inferred by the progressively updating decision boundaries (detailed later), our instance discrimination (ID) objective in terms of $\boldsymbol{I}_i$ is to match $\boldsymbol{x}_i$ with $\hat{\boldsymbol{x}}_i$ against its contrastive set $\tilde{X}_i = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \cdots, \tilde{\boldsymbol{x}}_k\}\ s.t.\ \tilde{y}_i \neq \tilde{y}_j, \forall j \in [1, k]$ composed by k stale representations of its pseudo-negative samples:

$$\mathcal{L}_{\mathrm{ID}}(\boldsymbol{I}_i) = -\log \frac{\exp(\cos(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_i)/\tau)}{\sum_{\tilde{\boldsymbol{x}} \in \tilde{X}_i \cup \{\hat{\boldsymbol{x}}_i\}} \exp(\cos(\boldsymbol{x}_i, \tilde{\boldsymbol{x}})/\tau)} \tag{4.13}$$

where $\cos(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is the cosine similarity between a pair of representations and τ is the temperature to control the concentration degree of distribution. As the samples in the same clusters share a common contrastive set, they are indirectly pushed closer in the feature space regardless of any intra-cluster variations. Therefore, the learned visual features are sensitive to both intra-cluster visual discrepancies and inter-cluster similarities. This capability is absent to instance contrastive learning, in which intra- and inter-cluster variations are considered equal. Hence, our learned feature representation is geared towards being sensitive to cluster-wise visual characteristics.

Semantic memory. For cross-cluster instance discrimination, we introduce a novel semantic memory. It serves two functions: (1) playing the role of conventional memory bank to store a list of k representations for contrastive learning; (2) imposing cluster structures into training data.

Specifically, we manage C independent memory banks $\mathcal{M} = \{M_1, M_2, \cdots M_C\}$ each corresponding to one cluster with a size of $k/(C-1)$. For an image $\boldsymbol{I}_i$ with pseudo label $\tilde{y}_i$, we construct a contrastive set $\tilde{X}_i$:

$$\tilde{X}_i = \{\tilde{\boldsymbol{x}} | \tilde{\boldsymbol{x}} \in M_j \; \forall j \in [1, C] \text{ and } j \neq \tilde{y}_i\}. \tag{4.14}$$

As formulated in Eq. (4.14), there is always one memory bank left out for each sample and the rest Ms are concatenated as its contrastive set $\tilde{X}_i$ approximately in size k (rounding error) to support cluster discriminative feature representation learning. Specially, the size $(k/(C-1))$ of each independent memory bank M_i is deliberately designed to ensure the consistent size of contrastive sets in SCL and the baseline instance contrastive learning model (He et al. [69]), so as to get rid of its impacts in comparison experiments. For memory update, after every backward pass, the representation $\hat{\boldsymbol{x}}_i$ enqueues to $M_{\tilde{y}_i}$ with the oldest one inside removed.

Cluster discrimination. To discover the underlying concepts with unique visual characteristics, we infer their decision boundaries by reducing the visual redundancy among clusters, namely maximising the visual similarity of samples within the same clusters and minimising that between clusters (Figure. 4.2 (c)). Concretely, as the representation of samples with different pseudo labels are stored independently in the semantic memory bank, they can be taken as anchors to describe their corresponding clusters. Given a training sample $\boldsymbol{x}_i$, the semantic memory bank naturally serves as a classifier based on pairwise similarity:

$$\tilde{p}_{i,j} = \frac{\sum_{\tilde{\boldsymbol{x}} \in M_j} \exp(\cos(\boldsymbol{x}_i, \tilde{\boldsymbol{x}})/\tau)}{\sum_{j'=1}^{C} \sum_{\tilde{\boldsymbol{x}} \in M_{j'}} \exp(\cos(\boldsymbol{x}_i, \tilde{\boldsymbol{x}})/\tau)}, \tag{4.15}$$

where $\tilde{p}_{i,j}$ implies the estimated probability that samples $\boldsymbol{x}_i$ should be assigned to the j-th cluster. With such potential memberships determined by sample-anchor visual similarities, we formulate a consistency loss for cluster discrimination (CD):

$$\boldsymbol{p}_i = \text{Softmax}(W^\top \boldsymbol{x}_i + B) \; \in \mathcal{R}^C, \tag{4.16}$$

$$\mathcal{L}_{\text{CD}} = \frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{C} -\tilde{p}_{i,j} \log p_{i,j}, \tag{4.17}$$

where $\{W; B\}$ is the learnable parameters of classifier $f_{\boldsymbol{\phi}}$ and n_{bs} denotes the size of mini-batch. In Eq. (4.17), we aim to minimise the cross-entropy of the distance-based cluster assignments

$\tilde{\boldsymbol{p}}_i$ and the predictions $\boldsymbol{p}_i$ yielded by the cluster decision boundaries and propagate the gradient back to $\boldsymbol{p}_i$ only to avoid feature learning from unreliable boundaries. By doing so, samples are assigned to the cluster with the most similar anchors while each cluster holding its own visual characteristics that make it different from others and correspond to an underlying semantic class with consistent and unique visual characteristics.

With the updated classifier $f_{\boldsymbol{\phi}}$, we renew the cluster assignments in a maximum likelihood manner (Eq. 4.1) for semantic memory construction in Eq. (4.14). As the predictions become increasingly more accurate along training, this update encourages the visual features derived by cross-cluster instance discrimination to be more aware of cluster-wise visual information.

Remarks. To avoid the learning process collapsing to extremely imbalanced cluster assignments, we adopt the "merge-and-split" strategy (Zhan et al. [215]) to update cluster assignments. This helps stabilise training. The process merges iteratively each over-small cluster with the largest cluster, then splits the merged into two new partitions, and repeats until there is no imbalanced cluster.

Hard samples mining. To enhance discrimination capacity, we identify semantically ambiguous samples and emphasise them in instance discrimination:

$$s_i^e = s_i^{e-1} + \mathbb{1}[\tilde{y}_i^e \neq \tilde{y}_i^{e-1}], \quad w_i^e = \frac{s_i^e}{\sum_{j=1}^{n_{\text{bs}}} s_j^e}, \tag{4.18}$$

$$\mathcal{L}_{\text{ID}} = \sum_{i=1}^{n_{\text{bs}}} w_i^e \mathcal{L}_{\text{ID}}(\boldsymbol{I}_i), \tag{4.19}$$

where w_i^e is the weights of $\boldsymbol{I}_i$ at the e-th training epoch. The samples that are frequently swapped across clusters (*i.e.*, hard samples) are assigned with higher weights for offering more useful discriminative learning clues.

4.2.3 Model Training

Given the instance (Eq. (4.19)) and cluster (Eq. (4.17)) discrimination losses, the overall training objective of SCL is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{ID}} + \lambda_2 \mathcal{L}_{\text{CD}}. \tag{4.20}$$

In general, the two weight parameters λ_1 and λ_2 in Eq. (4.20) are to balance the significance of $\mathcal{L}_{\text{ID}}$ and $\mathcal{L}_{\text{CD}}$. However, in the absence of labelled validation data in unsupervised clustering, here we set both weights to 1 to avoid exhaustive per-dataset parameter tunning. The model is trained by minimising $\mathcal{L}$. To that end, the weights of encoder $\boldsymbol{\theta}$ as well as the decision boundaries $\boldsymbol{\phi}$ are

updated by back-propagation and the momentum encoder $\hat{\boldsymbol{\theta}}$ is by $\hat{\boldsymbol{\theta}} \leftarrow \eta\hat{\boldsymbol{\theta}} + (1-\eta)\boldsymbol{\theta}$ where η is a momentum coefficient (He et al. [69]). Both objective functions (Eq. (4.19) and Eq. (4.17)) are differentiable thus can be trained end-to-end by the conventional stochastic gradient descent algorithm. The overall training procedure of SCL is summarised in Algorithm 4.

Algorithm 4: SCL for deep clustering.

Input: Unlabelled data $\mathcal{I}$, training epochs N_{ep}, iterations per epoch N_{it}, target cluster number C.

Output: A deep clustering model.

for epoch $= 1$ **to** N_{ep} **do**

  for iter $= 1$ **to** N_{it} **do**

    Generating a random mini-batch of unlabelled data;

    Generating two perturbed copies of the mini-batch;

    Computing $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ by $f_{\boldsymbol{\theta}}$ and $f_{\hat{\boldsymbol{\theta}}}$, respectively;

    Constructing sample-wise contrastive sets $\tilde{X}$ (Eq. (4.14));

    Computing the instance discrimination loss (Eq. (4.19));

    Constructing distance-based predictions $\tilde{\boldsymbol{p}}$ (Eq. (4.15));

    Feeding $\boldsymbol{x}$ into $f_{\boldsymbol{\phi}}$ to compute $\boldsymbol{p}$ (Eq. (4.16));

    Computing the cluster discrimination loss (Eq. (4.17));

    Updating $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by back-propagation;

    Updating $\hat{\boldsymbol{\theta}}$ by momentum;

    Updating the semantic memory by $\hat{\boldsymbol{x}}$;

  end

  Updating assignments (Eq. (4.1));

  Updating sample's weights (Eq. (4.18));

end

4.3 Experiments and Evaluations

4.3.1 Datasets and Metrics

Datasets. Evaluations were conducted on six challenging object recognition benchmarks. **(1) CIFAR-10/CIFAR-100 (Krizhevsky and Hinton [92])**: Natural image datasets composed by

60,000 samples in size 32×32 that are uniformly drawn from 10(/100) classes. **(2) STL-10 (Coates et al. [33])**: An ImageNet adapted dataset consists of 1,300 images from each of 10 classes in size 96×96. Additional 100,000 samples from unknown classes with no overlap with the known categories on training or test sets were deprecated in our experiments. **(3) ImageNet-10/Dogs (Russakovsky et al. [154])**: ImageNet subsets containing samples from 10 randomly selected classes or 15 dog breeds. **(4) Tiny-ImageNet (Le and Yang [97])**: Another ImageNet subset at large scale with 100,000 samples in size 64×64 evenly distributed in 200 classes. All the models were trained on the same data in consistent sizes as (Ji et al. [85]) for fair comparisons. We adopted the clustering setup same as (Ji et al. [85]; Wu et al. [193]; Chang et al. [20]): Using both the training and test sets (without labels) for CIFAR-10/CIFAR-100 and STL-10, while only the training set for ImageNet-10, ImageNet-Dogs and Tiny-ImageNet; The 20 super-classes on CIFAR-100 were considered as ground-truth.

Evaluation metrics. We used three standard clustering performance metrics: (a) Accuracy (**ACC**) is computed by assigning each cluster with the dominating class label and taking the average correct classification rate as the final score, (b) Normalised Mutual Information (**NMI**) quantifies the normalised mutual dependence between the predicted labels and the ground-truth, and (c) Adjusted Rand Index (**ARI**) evaluates the clustering result as a series of decisions and measures its quality according to how many positive/negative sample pairs are correctly assigned to the same/different clusters. All of these metrics scale from 0 to 1 and higher values indicate better performance.

4.3.2 Implementation Details

For fair comparisons, we followed the same practices of (Ji et al. [85]) to adopt a variant of ResNet-34 for small inputs as the backbone network. Given the different baseline models adopted in PICA and SCL, their own implementation are provided separately as below.

Partition confidence maximisation. Following (Ji et al. [85]), we used the auxiliary over-clustering strategy in a separate clustering head to exploit the additional data from irrelevant classes if available. For the over-clustering head, we set 700 clusters for Tiny-ImageNet (due to more ground-truth classes) and 70 clusters for all the others. The over-clustering head, discarded finally in test, was trained alternatively with the primary head. In case of no auxiliary data, we used the target data in over-clustering head, which plays a role of auxiliary learning. For training,

we used Adam optimiser (Kingma and Ba [89]) with a fixed learning rate 0.0001. All the models were randomly initialised and trained with 200 epochs. The regularisation penalties to model weights were set to be 0. We set the batch size to 256 and repeated each in-batch sample 3 times. Three operations, including random rescale, horizontal flip and colour jitters, were adopted for data perturbations and augmentation. We applied the sobel filter for restraining the model from capturing meaningless patterns of trivial colour cues. The weight of entropy regularisation in Eq. (4.12) was set to 2 empirically.

Semantic contrastive learning. For SCL training, we followed most the implementation choices of (Chen et al. [29]). All the models and the cluster assignments are randomly initialised. An SGD optimiser was adopted for model updates with weight decay in $5e-4$. The coefficient for momentum encoder updating was set to 0.9 and τ in Eq. (4.13) was 0.1. We stored $4096/(C-1)$ representations for each cluster in the semantic memory (Eq. (4.14)) on all the datasets except for $8192/(C-1)$ on Tiny-ImageNet due to larger scale. The learning rate was set to 0.03 with a batch size of 256 and cosine schedule (Loshchilov and Hutter [112]) was used for learning rate adjustments across 200 training epochs. Besides the target 'clustering' tasks which partition the target data into the ground-truth number of clusters to facilitate comparisons with previous works, we followed (Ji et al. [85]) to jointly train our model with auxiliary 'under-clustering' and 'over-clustering' tasks so to explore multi-grained visual similarity. The cluster number in 'under-clustering' was half of the ground-truth while sample specificity learning was considered as extreme 'over-clustering'. At test time, we used the assignments yielded by the classifier for 'clustering' tasks while the other two classifiers were deprecated.

Regardless the different implementations of the two models, we used the same hyper-parameters in each individual for all the experiments on different benchmark datasets *without* exhaustive per-dataset tuning which is unscalable, inconvenient nor unfriendly in deployment.

4.3.3 Comparisons with the State-of-the-Art

Deep Clustering. Table 4.1 compares the proposed PICA and SCL models with 13 state-of-the-art deep clustering models including both without contrastive learning (DEC (Xie et al. [200]), DAC (Chang et al. [20]), ADC (Haeusser et al. [63]), DDC (Chang et al. [21]), DCCM (Wu et al. [193]), IIC (Ji et al. [85]), DCCS (Zhao et al. [226]), GAT (Niu et al. [130])) and with instance contrastive learning (SCAN (Van Gansbeke et al. [181]), IDFD (Tao et al. [175]), CC (Li

Model	Instance	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs			Tiny-ImageNet		
		NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	✗	0.087	0.229	0.049	0.084	0.130	0.028	0.125	0.192	0.061	0.119	0.241	0.057	0.055	0.105	0.020	0.065	0.025	0.005
DEC*	✗	0.257	0.301	0.161	0.136	0.185	0.050	0.276	0.359	0.186	0.282	0.381	0.293	0.122	0.195	0.079	0.115	0.037	0.007
DAC*	✗	0.396	0.522	0.306	0.185	0.238	0.088	0.366	0.470	0.257	0.394	0.527	0.302	0.219	0.275	0.111	0.190	0.066	0.017
ADC*	✗	-	0.325	-	-	0.160	-	-	0.530	-	-	-	-	-	-	-	-	-	-
DDC*	✗	0.424	0.524	0.329	-	-	-	0.371	0.489	0.267	0.433	0.577	0.345	-	-	-	-	-	-
DCCM*	✗	0.496	0.623	0.408	0.285	0.327	0.173	0.376	0.482	0.262	0.608	0.710	0.555	0.321	0.383	0.182	0.224	0.108	0.038
IIC	✗	-	0.617	-	-	0.257	-	-	0.610	-	-	-	-	-	-	-	-	-	-
DCCS*	✗	0.569	0.656	0.469	-	-	-	0.376	0.482	0.262	0.608	0.710	0.555	-	-	-	-	-	-
GAT*	✗	0.475	0.610	0.402	0.215	0.281	0.116	0.446	0.583	0.363	0.594	0.739	0.552	0.281	0.322	0.163	-	-	-
PICA	✗	0.591	0.696	0.512	0.310	0.337	0.171	0.611	0.713	0.531	0.802	0.870	0.761	0.352	0.352	0.201	0.277	0.098	0.040
SCAN	✓	0.712	**0.818**	0.665	0.441	0.422	0.267	0.654	0.755	0.590	-	-	-	-	-	-	-	-	-
IDFD	✓	0.711	0.815	0.663	0.426	0.425	0.264	0.643	0.756	0.575	**0.898**	**0.954**	**0.901**	0.546	0.591	0.413	-	-	-
CC	✓	0.705	0.790	0.637	0.431	0.429	0.266	**0.764**	**0.850**	**0.726**	0.859	0.893	0.822	0.445	0.429	0.274	0.340	0.140	0.071
CRLC	✓	0.679	0.799	0.634	0.416	0.425	0.263	0.729	0.818	0.682	0.831	0.854	0.759	0.461	0.484	0.297	-	-	-
SCL	✓	**0.744**	0.813	**0.683**	**0.477**	**0.482**	**0.314**	0.593	0.638	0.485	0.877	0.930	0.861	**0.728**	**0.763**	**0.652**	**0.337**	**0.172**	**0.080**

Table 4.1: Comparisons with the state-of-the-art methods in deep clustering. The column 'Instance' indicate whether or not the approaches conducted deep clustering by instance contrastive learning. Methods with $(\cdot)^*$ trained without the additional data on STL-10. The $1^{st}/2^{nd}$ best results are indicated in **red**/blue.

et al. [106]), CRLC (Do et al. [39])). We made the following observations:

- Both the PICA and SCL models yielded superior performances on all benchmarks when being compared with contemporary approaches which are formulated without benefitting from instance contrastive learning (upper part of Table 4.1). Specifically, PICA surpasses all the strong competitors in most cases, sometimes by a large margin. Taking ACC for example, PICA outperforms the second-best competitor on CIFAR-10 and ImageNet-10 with 4.0% and 13.1% respectively while the performance gain over the baseline model (Ji et al. [85]) on STL-10 is 10.3%. Besides, the overall performance margins obtained by all instance contrastive learning based deep clustering methods (lower part of Table 4.1), demonstrate compellingly the benefit of contrastive constraints in unsupervised semantic concepts inference learning. Falling in this group of methods, our SCL model yielded also competitive performances on all benchmarks, outperforming the second-best model in four out of the six datasets, sometimes by a significant margin in some cases.

- For the methods built without relying on the visual features yielded by instance contrastive learning, DCCM (Wu et al. [193]) serves as a strong competitor on most datasets except for STL-10, on which it is outperformed by both IIC (Ji et al. [85]) and our PICA by more significant margins. We attribute this to the capability of exploiting auxiliary data of both

winner methods. Also, PICA is clearly superior to IIC for clustering the images of STL-10, suggesting the outstanding potential of our method for capitalising extra data during deep clustering. However, the auxiliary data ($10\times$) on STL-10 are sampled from the same distribution as the target data but guaranteed from different classes. Those additional data are beneficial to learn better representations but were explicitly excluded when training the target classifier (Li et al. [106]; Do et al. [39]). For the experiments of SCL model, we avoided using them because it is less practical to have such similar and guaranteed negative data unless their class labels are available. In this case, SCL's performance advantages over the methods learned without additional data are still notable (improved (Niu et al. [130]) by 5.5% w.r.t. clustering accuracy).

- The advantages obtained by PICA over contemporary methods on the more challenging ImageNet-Dogs (fine-grained) and Tiny-ImageNet (large-scale) benchmarks are relatively smaller. This is not surprised, since these datasets present higher inter-class similarity or complexer intra-class variations. By conducting instance-level image discrimination when learning visual features, the SCL model yielded its unique superiority in these cases, surpassing the second-best competitor (IDFD (Tao et al. [175]) and CC (Li et al. [106])) on ImageNet-Dogs and Tiny-ImageNet with 29.1% and 22.9% relative margins.

- Compared to the concurrent attempts on deep clustering by instance contrastive learning (Van Gansbeke et al. [181]; Tao et al. [175]; Li et al. [106]; Do et al. [39]), the SCL's superior performances is due to solving the contradiction between optimising instance contrastive learning (pull apart) and maximising intra-cluster compactness (push closer) by satisfying a consistency condition jointly by both objectives.

Representation Learning. Beyond the methods intrinsically designed for clustering (Van Gansbeke et al. [181]; Tao et al. [175]; Li et al. [106]; Do et al. [39]), we also compared the presented SCL model with a clustering-based representation learning approach (Caron et al. [15]) and two general instance contrastive learning schemes: Instance-wise learning (MoCo (Chen et al. [29])) and local neighbourhood discrimination based learning (PAD introduced in Chapter 3). The learned feature representations from both models are applied with k-means for clustering. As shown in Table 4.2 (top), our SCL method outperformed all the representation learning methods across the board. This shows clearly the advantages of SCL from holistically modelling the in-

Model	CIFAR-10	CIFAR-100	STL-10
	Clustering (ACC)		
MoCo	0.528	0.360	0.561
PAD	0.626	0.288	0.465
DeepCluster	0.374	0.189	0.334
SCL	**0.813**	**0.482**	**0.638**
	*k*NN classification		
MoCo	0.853	0.713	**0.772**
SCL	**0.871**	**0.721**	0.712

Table 4.2: Comparisons to representation learning methods in deep clustering. Results of MoCo were reproduced from scratch using the authors' code (Chen et al. [29]) and that of DeepCluster are from (Ji et al. [85]).

herent class structure, resulting in also a more optimal representation, as compared to separating representation learning from class membership estimation.

In addition to clustering, we further evaluated the generalisation ability of the image features derived by our SCL model and that of MoCo (He et al. [69]; Chen et al. [29]) in Table 4.2 (bottom). Specifically, we followed (Wu et al. [195]; He et al. [69]) to classify the *unseen* test images according to their k-nearest neighbours in the training set with a ResNet18 as backbone. Our competitive performance demonstrates the potential of learning generalisable visual features by exploring inherent class memberships without manual labels. Figure 4.3 further shows visual examples of cluster assignments made by instance contrastive learning (MoCo) and SCL, together with their corresponding confidence scores. It is evident that the class decision boundaries estimated by instance contrastive learning are not as separable as those by SCL, with the former giving nearly uniform predictions in many cases similar to the two examples shown. Moreover, such partitions became even less accurate when there are visual overlaps between semantic concepts, common in practice *e.g.* airship and ship. In contrast, SCL yielded consistently reliable and confident predictions by optimising jointly cross-cluster instance contrastive property and cluster compactness condition. This demonstrates the benefit of modelling explicitly underlying global class structures, which is the essence of SCL.

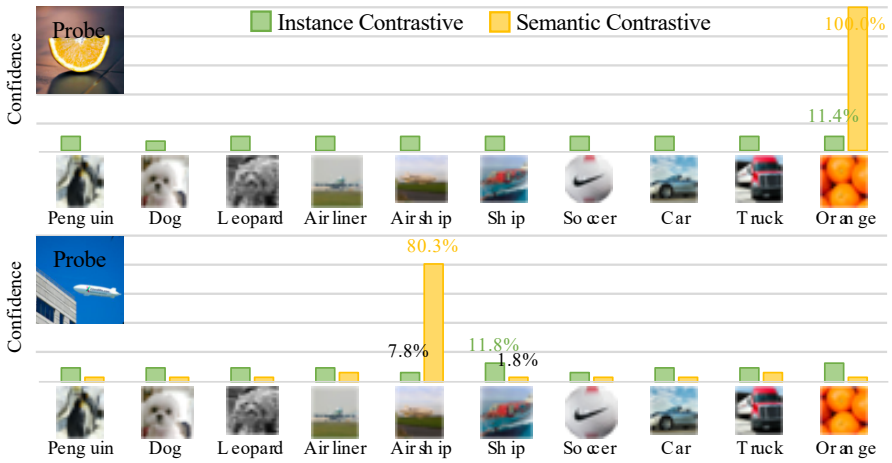


Figure 4.3: Visualisations of sample's cluster assignment and confidence scores yielded by Instance *vs*. Semantic contrastive learning on ImageNet-10.

4.3.4 Component Analysis of PICA

We conducted ablation studies to investigate the effect of different design choices in PICA.

Partition confidence dynamics. We started with examining the clustering confidence dynamics during training, which underpins the key idea of our PICA. In this examination, we used the maximum prediction probability (Eq. (4.1)) of every image to measure the clustering confidence, and summarised their 50-bins histogram statistics. We performed this test on CIFAR-10 at four accuracy performance milestones: 0.10 (random guess), 0.30, 0.50 and 0.70. As shown in Fig-

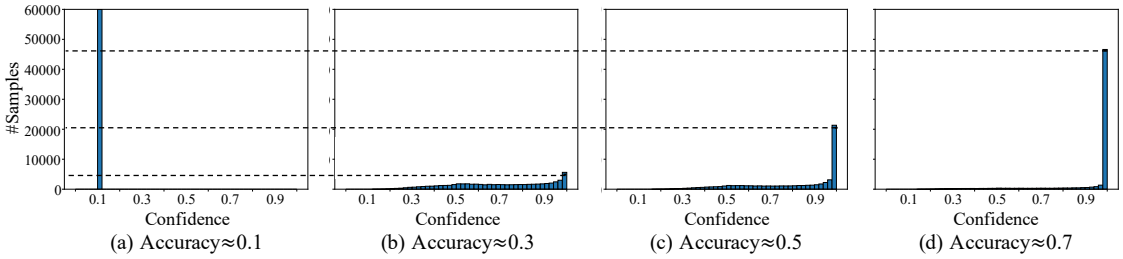


Figure 4.4: Partition confidence evolution of PICA in training on CIFAR-10. The probability range $[0, 1]$ is uniformly divided into 50 bins and the number of samples whose maximal prediction value is falling in each interval is counted.

ure 4.4, **(a)** the model started with random clustering close to a uniform prediction; **(b,c)** Along with the training process, an increasing number of samples get more confident cluster assignment; **(d)** At the end of training, a majority of samples can be assigned into clusters with 0.98+ probability confidence, nearly one-hot predictions.

Avoiding under-clustering. We examined how important PICA needs to solve the generic "under-clustering" problem (assigning most samples to a few clusters, *i.e.*, trivial solutions) in our context. The results in Table 4.3 indicate that it is highly necessary to take into account this problem in model design otherwise the model will be trivially guided to such undesired solutions. This also verifies that the proposed PICA idea is compatible well with the entropy regularisation of the cluster size distribution (Eq. (4.11)), enabling to eliminate simply trivial results without resorting to complex designs or tricks.

Entropy	CIFAR-10	CIFAR-100	ImageNet-10
✗	0.246	0.168	0.650
✓	**0.696**	**0.330**	**0.829**

Table 4.3: Effects of avoiding under-clustering using the entropy regularisation of the cluster size distribution. Clustering accuracies are reported.

Effect of over-clustering. We examined the performance contribution of over-clustering in PICA which serves two purposes: (1) Leveraging extra auxiliary data from irrelevant classes for mining more information (*e.g.* on STL-10); (2) In case of no auxiliary data, playing a role of ensemble learning (*e.g.* on CIFAR-10). The results are given in Table 4.4. It is clear that over-clustering helps in both cases, and interestingly the margin on CIFAR-10 is even bigger than that on STL-10. Also note that without using over-clustering, our PICA can still achieve competitive performances (*cf*. Table 4.1).

Over-clustering	CIFAR-10	STL-10
✗	0.582	0.633
✓	**0.696**	**0.687**

Table 4.4: Effects of over-clustering in PICA. STL-10 has auxiliary data whilst CIFAR-10 does not. Clustering accuracies are reported for comparison.

Robustness to data perturbation. We tested the clustering robustness against image transformations used for perturbing data distributions in PICA. Unlike existing methods typically using data augmentation by random perturbation at the local sample level, we exploit it at the global clustering solution level. The results in Table 4.5 show that our PICA requires data augmentation for offering strong performances. While seemingly surprised, this is also rational/sensitive since

our method performs the robustness enhance in a solution-wise manner; This effectively accumulate the augmentation effect of individual samples and potentially resulting in amplified effects eventually. However, this does not affect the use of PICA in general since data augmentation is just a standard necessary element of almost all deep learning methods.

Perturbation	CIFAR-10	CIFAR-100	ImageNet-10
✗	0.310	0.147	0.734
✓	**0.696**	**0.330**	**0.829**

Table 4.5: PICA's clustering robustness to data perturbation. The model trained without perturbations computes the matching matrix by the cosine similarities between the ASV determined by the original batch (without perturbation) only. Clustering accuracies are reported.

4.3.5 Component Analysis of SCL

Detailed ablation studies were conducted for in-depth analysis of SCL. k-means was adopted for models which did not yield desired number of clusters. Experimental results were averaged over multiple trials.

Instance and cluster discrimination. We investigated the independent contributions of our *cross-cluster* Instance Discrimination and *online* Cluster Discrimination designs in the SCL model. We took the MoCo as the baseline without both the ID and CD components. For models trained without cross-cluster ID, all the memory banks were concatenated as the contrastive set for every sample (Eq. (4.14)), whilst the cluster assignments $\tilde{\boldsymbol{p}}$ yielded by the semantic memory (Eq. (4.15)) was used for pseudo labels updating if learned without online CD. As shown in Figure 4.5, the models trained without cross-cluster ID or online CD can always surpass MoCo with remarkable margins, which demonstrates their effectiveness as individual components. By jointly learning with both, SCL always produced superior performances which indicates the mutual benefits of representation learning and decision boundaries reasoning.

Effects of cluster number. Jointly learning from multiple clustering tasks has been validated to be beneficial to deep clustering (Ji et al. [85]). Whilst there is no universal principle to determine the proper cluster number, we considered that the unlabelled images can be grouped in three different possibilities: fewer, equal or more clusters than the true number of classes (unknown). Specifically, we took instance-wise discrimination as the extreme case of 'over-clustering' and

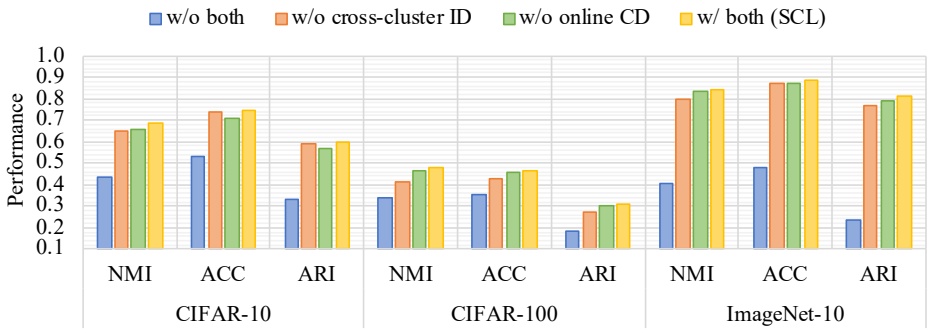


Figure 4.5: Ablation studies of *cross-cluster* Instance Discrimination and *online* Cluster Discrimination designs. Instance contrastive learning is trained with neither component.

halved the ground-truth numbers in 'under-clustering'. The SCL model was jointly learned with all these three tasks. Figure 4.6 shows that all the models trained with consideration of cluster assignments can better learn the ground-truth memberships than instance contrastive learning, which again indicates the defects of learning sample-specific representations on high-level semantic understanding of visual data. The SCL model explored multi-grained visual similarity which avoids it being misled by hard negative samples that are semantically different but visually similar in certain parts, and yielded the best results in all cases.

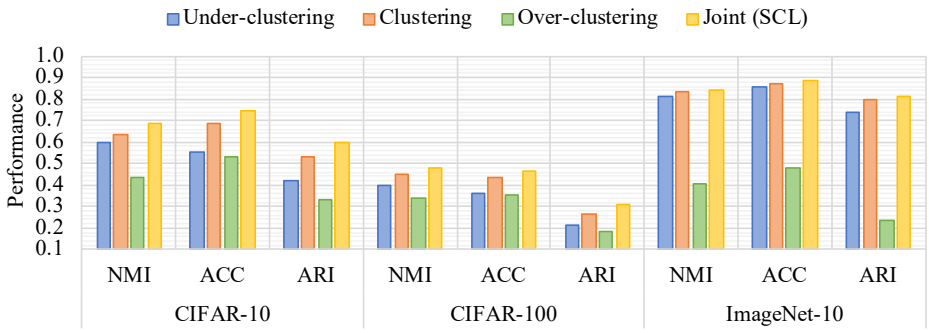


Figure 4.6: Effect of learning with different cluster numbers. The cluster numbers in over-, under- and clustering tasks are larger, smaller and equal to the ground-truth. SCL trains with all.

Hard sample mining. To emphasise the hard samples in model learning, we re-weighted the samples within the same mini-batches according to their assignment stability (Eq. (4.18)). To study the effectiveness of this design, we replaced it by averaging their losses as in conventional batch-wise training. According to Figure 4.7, the learned clusters show higher consistency with the ground-truth classes when training with the re-weighting strategy. This demonstrates the importance of paying more attention to hard samples with ambiguous semantic meanings so to further improve the model's class discrimination capability.

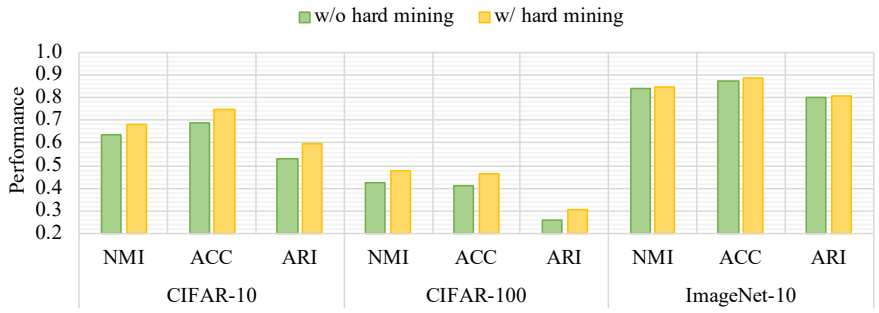
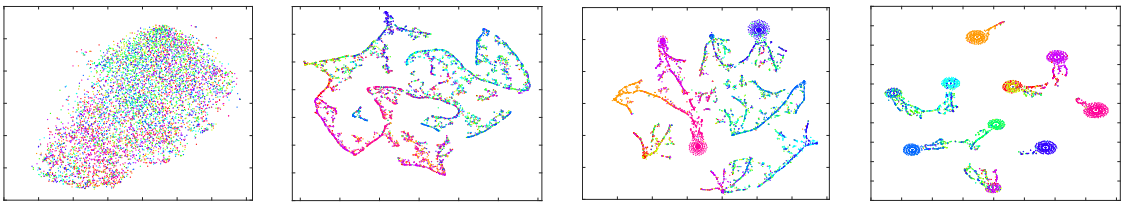


Figure 4.7: An ablation study on the hard sample mining strategy in SCL.

4.3.6 Visualisation and Qualitative Study

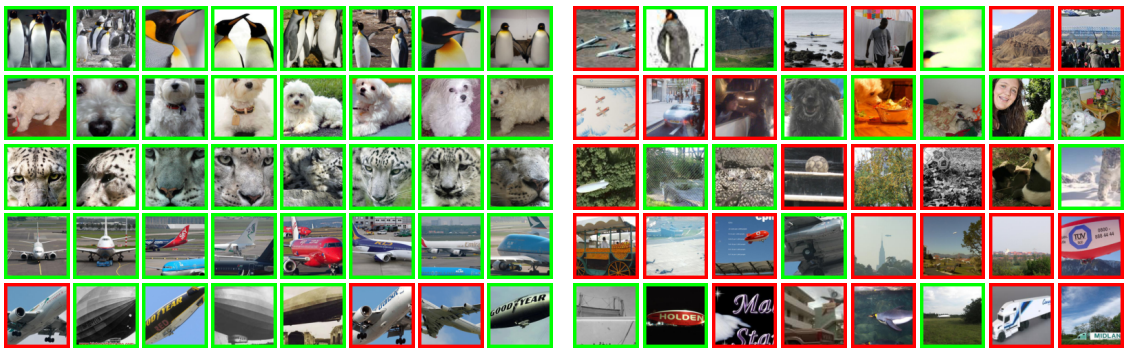
Evolution of cluster assignment. To provide a better understanding of how does the presented models work, we analysed the PICA model qualitatively by visualising the evolution of cluster assignment across the whole training process. This enables us to find out how does our model gradually attain the final result. We tracked the model status during the whole training process on CIFAR-10 and evaluated at four accuracy performance milestones: 0.10 (random guess), 0.30, 0.50 and 0.70. Using t-SNE (Maaten and Hinton [116]), we plotted the predictions of 6,000 randomly selected samples with the ground-truth classes colour encoded. Figure 4.8 shows that, **(a)** the model started from a chaotic status where all the samples were assigned into each cluster with similar probabilities; **(b)** With the supervision from the proposed objective, easy samples with most salient observations were separated gradually while the remaining were still indecisive; **(c)** As the training proceeded, easy samples served as references for the others and attracted those with high visual similarities; **(d)** Finally, the model converged to a stable clustering solution which separated samples from different classes with some confusion around decision boundaries.



(a) Accuracy=0.1 (b) Accuracy=0.3 (c) Accuracy=0.5 (d) Accuracy=0.7

Figure 4.8: Prediction dynamics of PICA across the training process on CIFAR-10. A total of 6,000 randomly sampled images were used. Ground-truth classes are colour coded.

Visual case examples. For qualitative study about SCL, we investigated the success and failure cases of cluster assignments predicted by it to offer extra insights into the model. Two types of cases were studied on ImageNet-10. Samples with the largest probability to be assigned into their corresponding cluster are shown in Figure 4.9 (a) cluster-wise 'confident' cases: From left-to-right, the left-most sample in the first row of Figure 4.9 (a) holds the highest probability to be assigned into that cluster (row) than others predicted to be in the same row (cluster). Figure 4.9 (b) shows the least confident samples of each cluster (row) from left-to-right. By comparing the 'confident' and 'unconfident' cases, it is evident that the assignment confidence yielded by the SCL model is closely aligned to the correctness of predictions, *i.e.*, samples with confident assignment are more likely to be assigned into the correct clusters. It means that the semantic concept encoded in our learned clusters are consistent with the ground-truth categories, which demonstrates the remarkable capability of the SCL model on exploring the high-level semantic meanings of unlabelled imagery data. Moreover, by examining the failed 'confident' cases, SCL suffers similarly as supervised learning classifiers such that it is prone to make incorrect predictions when the most discriminative elements (parts) of objects are missing, *e.g.* airliners were mistaken as airships when the empennage was invisible. From the examples of the 'unconfident' cases (but still succeeded in being clustered correctly), they demonstrate SCL's robustness to distractions even when only a small part of target objects are visible as long as they reveal sufficient discriminative information. Most of the failed cases share some common background visual characteristics away from the target objects. This suggests that it remains a challenging problem for unsupervised learning methods to always focus on the more salient and relevant content.



(a) Confident cases (b) Unconfident cases

Figure 4.9: Case-study examples of SCL from ImageNet-10. (a) Top-8 samples with the highest predicted probabilities to each class are shown in the order of more-to-less 'confident' cases from left-to-right, and (b) bottom-8 samples with the lowest probabilities (left-to-right) are 'unconfident' cases. Samples in green boxes are successfully assigned into the correct classes while those with red boxes are failed cases.

4.4 Summary

In this chapter, we proposed two new methods for jointly learning class discriminative visual representations and the global decision boundaries of clusters by addressing the two common challenges in deep clustering including lacking semantic plausibility and error-propagation from unreliable pseudo labels to feature learning. Specifically, we introduced a novel idea of learning the most promising and semantically plausible clustering solution from the partition confidence perspective, and formulated an elegant objective loss function based on a partition uncertainty index. The PICA model extends the idea of maximal margin clustering used in previous shallow models to the stronger deep learning paradigm with significant loss function formulation. It can be introduced in standard deep network models and end-to-end trainable without bells and whistles. Besides, a new Semantic Contrastive Learning method is presented which addresses the fundamental limitation of instance contrastive learning by imposing the cluster structure into the unlabelled training data so to jointly learn discriminative visual feature representations and reason about cluster decision boundaries while avoiding the inherent contradiction between their learning objectives. With the design of class-agnostic instance discrimination, our SCL model learns visual features with high robustness to temporal (intermediate) cluster assignments in the course of model training, which mitigates the common error-propagation problem of contemporary deep clustering techniques. Extensive experiments on six challenging datasets demonstrated the performance superiority of both the proposed methods (PICA and SCL) over a wide range of the state-of-the-art deep clustering as well as generic representation learning approaches on deriving class discriminative feature representations and their underlying semantic class memberships.

Chapter 5

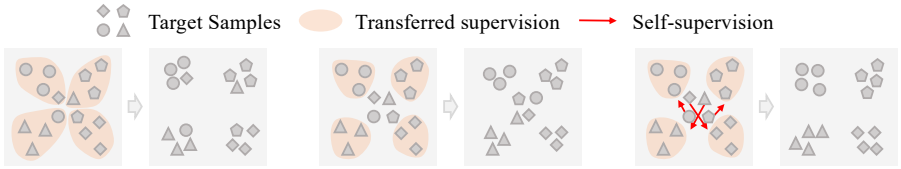
Transferring Knowledge from Relevant Labels with Self-supervised Remedy

---

This chapter investigates knowledge transfer from a relevant source data domain with manually defined category labels to an unlabelled target domain for visual object recognition, *i.e.* transfer clustering (Han et al. [65]). Given N^s image-label pairs $\{\boldsymbol{I}^s, y^s\}_{i=1}^{N^s}$ drawn from a source label space $\mathcal{Y}^s = \{1, 2, \cdots, C^s\}$ and N^t target images $\{\boldsymbol{I}^t\}_{i=1}^{N^t}$ from $\mathcal{Y}^t = \{1, 2, \cdots, C^t\}$ where $\mathcal{Y}^t \neq \mathcal{Y}^s$. In transfer clustering, no class label is annotated on target images. The objective is to jointly learn a feature representation of target samples $f_{\boldsymbol{\theta}} : \boldsymbol{I}^t \rightarrow \boldsymbol{x}^v$ and the probabilities that they belong to each of C^t clusters $f_{\boldsymbol{\phi}} : \boldsymbol{x}^v \rightarrow \boldsymbol{p}$ so that the target samples of the same class labels (unseen) are more likely to be allocated into the same partitions. Due to the absence of labelling in the target domain, the target distribution is agnostic, so as the discrepancy between it and the source distribution. It is therefore necessary to consider not only an effective way for source domain knowledge transfer in target sample clustering, but also how to identify and deal with those target samples having insufficient (weak or ambiguous) support from the source domain prior-knowledge. This is intrinsically challenging as the arbitrarily complex appearance patterns and variations exhibited in the imagery data usually lead to intricate relationships between source domain and target domain distributions.

This chapter is structured as follows: Section 5.1 provides an overview of the proposed method; Section 5.2 introduce the formulation of benefiting knowledge transfer by self-supervision in the cluster analysis; Section 5.3 elaborate the optimisation/training details of the model; Sec-

tion 5.4 exhibits the validation results by extensive experiments on both small and large-scale natural objects datasets and Section 5.5 summaries the chapter.



(a) Truncation (b) Selection (c) SUPREME

Figure 5.1: An illustration of strategies dealing with ambiguous transferred supervision. **(a)** Truncating ambiguous transferred supervision to make them determined. **(b)** Neglecting unconfident transferred supervision. **(c)** Complementing transferred supervision with self-supervision.

5.1 Approach Overview

For effective knowledge transfer across domains while mitigating the negative impacts of the non-transferable prior caused by distribution shift or class discrepancy, the proposed SUPREME method shares the spirit of auxiliary learning to train a model jointly by transferred knowledge and self-supervision constructed by the information intrinsically available in the unlabelled target domains. Comparing with existing knowledge transfer strategy, the SUPREME method holds its own advantages as below:

- The SUPREME model is more robust to unreliable transferred supervision. One straightforward solution to deal with ambiguous supervisions is "truncate" them to convert mandatorily to be determined (Figure 5.1 (a)), *e.g.* converting the assignment probabilities (soft-labels) decided by cluster analysis on the pre-learned feature space to pseudo hard labels according to the most confident assignments. However, this can be error-prone and rather arbitrary. The SUPREME model considers it is more consistent to replace those unreliable supervisions by the intrinsic information encoded in the target data, *i.e.* self-supervision (Figure 5.1 (c)). Although self-supervision is usually less complete than the prior-knowledge acquired from human annotations (Hsu et al. [79, 80]), it can minimise the misleading effect of applying non-transferable knowledge to the target domain.

- The SUPREME model is more informative when learning on hard target samples that yielding ambiguous transferred supervisions. Rather than truncating the uncertain supervisions, another intuitive idea is to selectively learn from the (easy) samples on which the transferred knowledge is applicable while neglecting the "hard" ones (Xie et al. [200]; Han

et al. [64]), which tends to introduce bias in model training and result in less discriminative feature space (Figure 5.1 (b)).

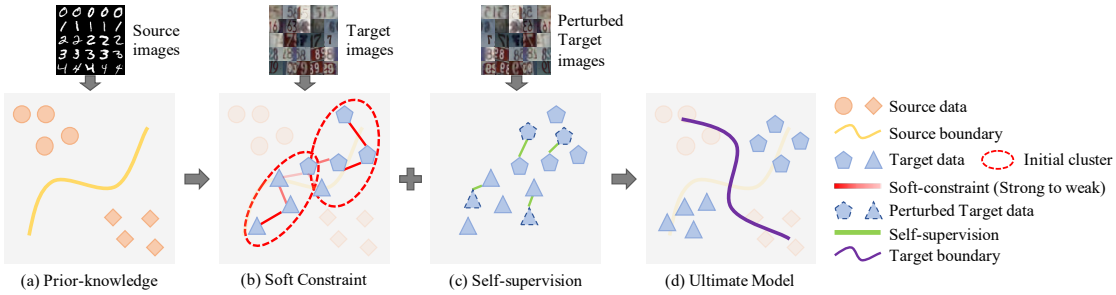


Figure 5.2: An overview of the proposed *self-SUPervised REMEdy* method.

An overview of the proposed self-SUPervised REMEdy model is shown in Figure 5.2. We first acquire the prior-knowledge from a source domain by learning a discriminative representational space to classify source images according to their manual category labels (Figure 5.2 (a)). k-means is then applied upon the resulted feature space to construct a *soft constraint* between each pair of target samples, whose confidence is measured by introducing a novel joint-entropy based metric (Figure 5.2 (b)). Self-supervision on model's invariance to image distortions is then introduced to make up for the ambiguous transferred supervisions (Figure 5.2 (c)). By jointly trained with both supervisions, our model finally learns a discriminative latent space and the decision boundaries for the target data (Figure 5.2 (d)).

5.2 Self-supervision in Transfer Clustering

We start from the construction of initial data partitions with the help of a model pretrained on a source domain for transferring prior-knowledge. Given a model $f_{\tilde{\boldsymbol{\theta}}}$ which is pretrained in supervised manner on source data $\{\boldsymbol{I}^s, y^s\}_{i=1}^{N^s}$, transfer learning assumes that there is some transferable common knowledge shared by the source and target domains, and some non-transferable specific knowledge to the source. To start, we yield an initial representation of target images $\boldsymbol{I}^t$ by feeding them into the pre-learned model: $\tilde{\boldsymbol{x}}^t = f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{I}^t)$. An initial clustering solution of $\tilde{\boldsymbol{x}}^t$ is computed by any standard technique (*e.g.* k-means). By separating $\tilde{\boldsymbol{x}}^t$ into C^t groups, we assumed that all the prior-knowledge of the source domain are applicable to the target domain by this initial clustering solution. This is not always true. To overcome this problem, we then formulate the transfer clustering as a constrained clustering task in which the constraints are formed by *pair-*

wise similarities between target samples determined by their initial assignment. The constraints are then weighted by their confidence estimated by the concentration degree (entropy) of their prior assignment probability distribution.

Construction of transferred supervision. Given the representations $\tilde{\boldsymbol{x}}^t$ and the C^t clusters centroid $\{\boldsymbol{c}_i\}_{i=1}^{C^t}$, we measure the probability $\tilde{\boldsymbol{p}}(\boldsymbol{c}_j|\boldsymbol{I}_i^t)$ that sample $\boldsymbol{I}_i^t$ belongs to cluster $\boldsymbol{c}_j$ by computing a student's t-distribution (Xie et al. [200]):

$$\tilde{\boldsymbol{p}}(\boldsymbol{c}_j|\boldsymbol{I}_i^t) = \frac{(1+\|\tilde{\boldsymbol{x}}_i^t - \boldsymbol{c}_j\|^2/\alpha)^{-(\alpha+1)/2}}{\sum_{k=1}^{C^t}(1+\|\tilde{\boldsymbol{x}}_i^t - \boldsymbol{c}_k\|^2/\alpha)^{-(\alpha+1)/2}} \tag{5.1}$$

The parameter α in Eq. (5.1) is the freedom of student's t-distribution and is set to 1 in our implementation. We denote $\tilde{\boldsymbol{p}}(\boldsymbol{c}_j|\boldsymbol{I}_i^t)$ by $\tilde{\boldsymbol{p}}_{i,j}$ for brevity. With the initial assignment probabilities, we then estimate how likely two target samples are from the same class by the inner product between their initial distributions:

$$\tilde{r}_{i,j} = \tilde{\boldsymbol{p}}_i^\top \cdot \tilde{\boldsymbol{p}}_j = \sum_{k=1}^{C^t} \tilde{p}_{i,k} \cdot \tilde{p}_{j,k} \tag{5.2}$$

The joint probabilities of sample pairs $\tilde{r}_{i,j}$ reaches its maximum $\tilde{r} \to 1$ only when two target samples are both close to the same cluster's centroid. In which case, they are considered "confidently positive". Otherwise, the pairwise relation is either ambiguous (neither samples is close to any cluster centroid) or "confidently negative" (two samples are close to different cluster centroids), resulting in the positive probability between them becomes the minimum $\tilde{r} \to 0$. To construct the transferred supervision signals, we then train a CNN model with soft constraints to encourage target pairs with large $\tilde{r}_{i,j}$ to be assigned into the same groups by a binary cross entropy loss (BCE):

$$\mathcal{L}_{\text{BCE}}(i,j) = \tilde{r}_{i,j}\log r_{i,j} + (1-\tilde{r}_{i,j})\log(1-r_{i,j}), \tag{5.3}$$

where n is the mini-batch size and $r_{i,j} = \sum_{k=1}^{C^t} p_{i,k} \cdot p_{j,k}$ is the up-to-date positive probability of the sample pair consisting of $\boldsymbol{I}_i^t$ and $\boldsymbol{I}_j^t$ and is computed by $\{\boldsymbol{p}_i, \boldsymbol{p}_j\} = f_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\{\boldsymbol{I}_i^t, \boldsymbol{I}_j^t\}))$.

Constraint weighting. In the formulation of $\mathcal{L}_{\text{BCE}}$ (Eq. (5.3)), ambiguous and confidently negative sample pairs hold similar low probabilities to be assigned into the same clusters. However, comparing with the latter, samples of ambiguous pairs are more likely positive. It means that the prior-knowledge transferred to the confidently negative pairs are more reliable than that to the ambiguous pairs. Instead of learning from all the prior in equal importance, the model should be encouraged to focus on the transferable knowledge. To that end, we assume that target samples

near the initial cluster centroids are able to form confident pairwise relations while those close to the decision boundaries cannot. Subsequently, we quantify the confidence of the pairwise relations by the joint entropy of initial assignment distributions:

$$\begin{aligned} H(\boldsymbol{I}_i^t, \boldsymbol{I}_j^t) &= -\sum_{k=1}^{C^t}\sum_{k'=1}^{C^t} \tilde{p}_{i,k}\tilde{p}_{j,k'} \log \tilde{p}_{i,k}\tilde{p}_{j,k'}, \\ H_{\max} &= \log(C^t)^2, \\ w_{i,j} &= \frac{\exp((H_{\max} - H(\boldsymbol{I}_i^t, \boldsymbol{I}_j^t)/(H_{\max}\cdot\tau))}{\sum_{i',j'}^{n}\exp((H_{\max} - H(\boldsymbol{I}_{i'}^t, \boldsymbol{I}_{j'}^t))/(H_{\max}\cdot\tau)}, \end{aligned} \tag{5.4}$$

where τ is the temperature that controls the concentration of the confidence distribution; $H(\boldsymbol{I}_i^t, \boldsymbol{I}_j^t)$ is the joint entropy of $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{p}}_j$; $w_{i,j}$ is the normalised confidence of constraint $\tilde{r}_{i,j}$. The overall penalty of a mini-batch will then be determined by the weighted sum instead of the average of Eq. (5.3):

$$\mathcal{L}_{\text{clu}} = -\sum_{i=1}^{n_{\text{bs}}}\sum_{j=1}^{n_{\text{bs}}} w_{i,j}\mathcal{L}_{\text{BCE}}(i,j). \tag{5.5}$$

The weighting strategy measures the reliability of prior-knowledge from the source domain in terms of different target samples, which mitigates the misleading effects caused by applying non-transferable prior to target domain.

Self-supervision information source. As determined by the cost function $\mathcal{L}_{\text{clu}}$ (Eq. (5.5)), samples falling into an ambiguous area of the initial feature space (*e.g.* those with similar distances to multiple cluster centroids) make necessarily less contribution to knowledge transfer as they are given smaller weights. On the other hand, those samples can play a significant role in learning a more discriminative feature space in the target domain. Due to the absence of ground-truth labels and the ineffectiveness of source domain prior-knowledge on these "hard" samples, the information we can leverage instead for additional supervision on model learning are intrinsic characteristics of those "hard" target images. Inspired by recent unsupervised learning ideas, we formulate our self-supervision signals based on the idea to maximise model's invariance to trivial visual distortions (Wu et al. [195]; Ji et al. [85]; He et al. [69]) to impose guaranteed positive inter-sample relationships in model training. By training the model to produce one-hot and identical predictions for each unlabelled images and its randomly perturbed copies, we encourage a determined clustering function which yields concentrated prediction distributions and is capable of finding the common that is invariant to spatio or non-material distortions so to model the intra-class discrepancy without manual label.

To form the self-supervised constraints to complement ambiguous transferred supervisions, we apply a random set of image transformations $\mathcal{T}(\cdot)$ on the original raw imagery data, then compute the positive probability $r_{i,j}$ in Eq. (5.5) according to the assignment distribution of sample $\boldsymbol{I}_i^t$ and that of $\mathcal{T}(\boldsymbol{I}_j^t)$: $r_{i,j} = \sum_{k=1}^{C^t} p_{i,k} \cdot \mathcal{T}(p_{j,k})$ where $\mathcal{T}(p_j) = f_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathcal{T}(\boldsymbol{I}_j^t)))$. We then set $\tilde{r}_{i,i} = w_{i,i} = 1 \forall i \in [1,n]$ so that "hard" samples with ambiguous transferred supervision (small w) will be supervised mostly by the self-supervision to produce persistent assignments for them and their transformed copies. In this way, the two supervisions are integrated harmoniously.

5.3 Model Training

Beyond the self-supervised knowledge alignment objective $\mathcal{L}_{\text{clu}}$ (Eq. (5.5)), our model is also trained with several regularisations to refrain from degenerated solutions. The overall training objective is thereby to minimise the weighted sum of $\mathcal{L}_{\text{clu}}$ and all the regularisation terms.

Avoiding extremely imbalanced clusters. The training objective of cluster analysis encourages the maximisation of intra-cluster compactness and inter-cluster diversity, hence, the model can possibly collapse by assigning all the samples into one single cluster. To refrain from this, we introduce a *balance regularisation* on cluster size:

$$\mathcal{L}_{\text{balance}} = \log C^t - \sum_{k=1}^{C^t} z_k \log z_k, \quad z_k = \frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} p_{i,k} \tag{5.6}$$

where z_k is the approximated size of the k-th cluster and the maximal entropy $\log C^t$ is added to ensure positive regularisation values. We train the model to minimise the negative entropy of the approximated cluster size distribution so as to avoid extremely imbalanced distributions.

Pseudo attributes representational space. Given that the target and source recognition tasks are composed of different classes, the feature distributions in the target domain can not be learned by aligning with the source distribution based on the shared label space as in UDA (Ganin and Lempitsky [49]). Therefore, several approaches have been proposed to bridge models across domains and label spaces by modelling attributes (Fu et al. [48]; Rastegari et al. [146]). These approaches share a similar assumption as zero-shot learning that recognition in different relevant domains should be performable in a common attribute space (Chang et al. [22]). Motivated by this, given the visual feature $\boldsymbol{x}^v$ of sample $\boldsymbol{I}$ from either domains, we project it to the latent factor space produced by a linear layer $f_{\boldsymbol{\omega}}$ and activate it by the non-linear *Sigmoid* function

$\boldsymbol{x}^a = \sigma(f_{\boldsymbol{\omega}}(\boldsymbol{x}^v))$. An element-wise *binary regularisation* is then applied on the factor space:

$$\mathcal{L}_{\text{attr}} = -\frac{1}{n_{\text{bs}} \times D} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{D} x_{i,j}^a \log x_{i,j}^a + (1 - x_{i,j}^a) \log(1 - x_{i,j}^a), \tag{5.7}$$

where D is the dimension of the attribute representation $\boldsymbol{x}_i^a \in \mathbb{R}^D$. The binary regularisation attains its minimum $\mathcal{L}_{\text{attr}} \to 0$ when $x_{i,j}^a \in \{0,1\} \; \forall \; j \in [1,D]$. Afterwards, $\boldsymbol{x}^a$ will be fed into the domain-specific classifier to predict the assignment distribution. The prediction of target samples is supervised by our proposed objective function (Eq. (5.5)) while that of source data is by the conventional cross-entropy loss with the provided labels on source domain:

$$\mathcal{L}_{\text{xent}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{C^s} \mathbb{1}[j = y_i^s] \log p_{i,j}. \tag{5.8}$$

The $\mathbb{1}[\cdot]$ denotes the indicator function which equals to 1 *iff* j is the ground-truth label y_i^s, otherwise 0. Although our ultimate goal is on target domain, jointly training with the classification task on source domain can also be taken as a *multitasks regularisation* to avoid learning trivial representation as well as the well-known catastrophic forgetting problem in transfer learning (Goodfellow et al. [59]).

5.4 Experiments and Evaluations

5.4.1 Datasets, Protocols and Metrics.

Datasets. Evaluations of the proposed SUPREME method were conducted on 10 benchmarks. **CIFAR-10 (/CIFAR-100)** (Krizhevsky and Hinton [92]): An imagery dataset containing 50,000/ 10,000 training and testing data drawn from 10(/100) classes uniformly. **SVHN** (Netzer et al. [125]): The Street View House Numbers dataset includes 73,257/26,032 train/test images lying in 10 digit classes $0 \sim 9$. **ImageNet** (Russakovsky et al. [154]): A large scale imagery dataset with over 1.2 million images from 1,000 classes. **CUB** (Welinder et al. [192]): Caltech-UCSD-Birds contains 11,788 images from 200 breeds of birds with 312 binary attributes annotations. **FLO** (Nilsback and Zisserman [129]): Oxford Flower dataset gathers images from 102 flower categories with each class consisting of between 40 and 258 instances. **SUN** (Patterson and Hays [136]): SUN Attribute is another common fine-grained datasets used in ZSL with 14,303 images included. **AwA2** (Xian et al. [196]): Animals with Attributes2 consists of 37,322 images of 50 animals classes with 85 numeric attributes for each class. **Mini-ImageNet** (Ravi and Larochelle [147]): An Imagenet adapted dataset containing 100 classes with 600 samples in each category.

Following (Ren et al. [149]), 64/16/20 classes are selected as base/validation/query set, respectively. We adopted the base set as the source domain and the query set as the target one. **Tiered-ImageNet** (Ren et al. [149]): Another subset of ImageNet which is larger than Mini-ImageNet and its categories are selected with hierarchical structure so that the base and query classes are disjointed semantically. The base set is composed by 351 classes from 20 super-categories while the validation set contains 97 classes and 160 for the query set. Similarly, the base and query set are taken as the source and target domain respectively under the transfer clustering setting. Data examples from each datasets are shown in Figure 5.3.

Figure 5.3: Examples of datasets used in transfer clustering.

Protocols. To transfer knowledge across domains in an unsupervised manner, we assumed that human annotations were only available on source domains and took the number of target classes as the only prior. We aimed to provide a generic solution to unsupervised transfer learning with fewer assumptions than most of the existing settings. To that end, in addition to comparing with transfer clustering techniques following the same setups as (Han et al. [64]), we further evaluated the effectiveness of SUPREME on FSL and ZSL benchmarks. Note, SUPREME did not utilise any word-vector embedding space knowledge on either the source or the target class labels nor sample-wise manual label on target domain as compared to the ZSL and FSL methods. We followed the settings as in (Xian et al. [198]) and (Ren et al. [149]) for ZSL and FSL, respectively. Moreover, there are two different training schemes commonly adopted in ZSL and FSL denoted as 'transductive' and 'inductive'. Their main difference is that additional unlabelled data in target domains is available and used for model training in the transductive scheme while not in the inductive one. We compared the proposed SUPREME model with contemporary methods

designed for both the schemes.

Evaluation metrics. We adopted two standard metrics in cluster analysis for evaluation: **(a)** Clustering Accuracy is determined by the percentage of target samples that are assigned into the cluster which is matched with the correct ground-truth class by the Hungarian algorithms (Kuhn [94]). **(b)** Normalised Mutual Information quantifies the normalised mutual dependence between the predicted assignments and the ground-truth memberships. Both of these metrics are falling within the range of $[0,1]$ and higher values indicate better performances. In FSL/ZSL tasks, we reported clustering accuracy to be compared with the classification accuracy yielded by the competitors, both of which reveal the model's discrimination ability and are within the same scale.

5.4.2 Implementation Details

We used the same network architectures as the ones adopted by (Han et al. [64]) as well as the corresponding model weights pretrained on source domains provided by them on the transfer clustering evaluation and the ImageNet pretrained ResNet101 (Chao et al. [23]) in ZSL to be consistent with (Xian et al. [198]). In FSL, we followed (Wang et al. [190]) to use ResNet12 as our backbone and adopted their provided pretrained model weights. The Adam algorithm (Kingma and Ba [89]) is adopted for model training with a fixed learning rate ($1e-3$). The image transformations used for data perturbation include random rescale and random horizontal flip, which are also adopted by (Han et al. [64]). All the main results are averaged over 10 runs while the ones from ImageNet are averaged over 3 runs with different data splits following (Han et al. [64]; Hsu et al. [79, 80]).

5.4.3 Comparisons with the State-of-the-Art

Transfer Clustering. We first evaluated SUPREME's effectiveness on clustering unlabelled data with the help of prior-knowledge from source domains by comparing with seven state-of-the-art transfer clustering models including LPNMF (Cai et al. [14]), LSC (Chen and Cai [28]), KCL (Hsu et al. [79]), MCL (Hsu et al. [80]), DEC (Xie et al. [200]), DTC (Han et al. [64]) and AutoNovel (Han et al. [65]) as well as the baseline which applies k-means on the feature space pre-learned in source domains. Results in Table 5.1 show that: **(1)** Most of the unsupervised transfer clustering methods yielded superior performances than the k-means baseline. As a cross-domain deployment solution, k-means generally applies the knowledge acquired from one

Method	CIFAR-10		CIFAR-100		SVHN		ImageNet	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.655	0.422	0.622	0.555	0.426	0.182	0.719	0.713
LPNMF	-	-	-	-	-	-	0.430	0.526
LSC	-	-	-	-	-	-	0.733	0.733
KCL	0.665	0.438	0.274	0.151	0.214	0.001	0.738	0.750
MCL	0.642	0.398	0.327	0.202	0.386	0.138	0.744	0.762
DEC	0.749	0.572	0.721	0.630	0.576	0.348	0.784	0.790
DTC	0.875	0.735	0.728	0.634	0.609	0.419	0.784	**0.791**
AutoNovel*	0.917	-	0.752	-	**0.952**	-	**0.825**	-
SUPREME	0.914	0.794	0.758	0.663	0.808	0.606	0.795	0.778
SUPREME*	**0.920**	**0.810**	**0.794**	**0.716**	0.864	**0.715**	0.795	0.778

Table 5.1: Comparisons with the state-of-the-art methods in transfer clustering. The 1st/2nd best results were marked in **red**/blue. Methods with $(\cdot)^*$ used self-supervised representation learning for fine-tuning with the unlabelled target data on CIFAR-10, CIFAR-100 and SVHN. Results of previous methods were adopted from (Han et al. [64, 65]).

domain to another without any selection or adaptation. Its disadvantages demonstrate the necessity of dealing with distribution shift/discrepancy. **(2)** The proposed SUPREME method was on par with all the competitors on most datasets. This suggests the effectiveness of our auxiliary learning design, which leverages the self-supervision to fill in the gap where prior-knowledge is not transferable. **(3)** The advantages of SUPREME on ImageNet was weaker than that on others. While seemingly surprised, this is also rational due to the between/within classes variations are modelled by geometry transformations or colour perturbations applied to input data, which is intrinsically restricted comparing with the complex and diverse image variations exhibited in such a larger scale dataset. Nevertheless, this doesn't degrade the contribution of our key idea of complementing source domain transferred knowledge with target domain self-supervision. It encourages exploiting more fully on larger scale data. **(4)** As an ad-hoc training strategy, self-supervised pretraining with unlabelled data from both domains brought significant help to the knowledge transfer. This is credited to the narrowed gap between the distribution learned by pretrained model and that of the target domain.

Protocol	Method	Mini-ImageNet		Tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
Inductive	TapNet	0.617	0.764	0.631	0.803
	CTM†	0.641	0.805	0.684	0.843
	MetaOpNet	0.641	0.800	0.658	0.818
Transductive	TPN	0.595	0.757	0.587	0.743
	TEAM*	0.601	0.759	-	-
	LR+ICI	0.668	0.793	0.808	0.879
	SVM+ICI	0.658	0.789	0.806	0.879
	SUPREME	0.853		0.803	

Table 5.2: Comparisons with the state-of-the-art methods in few-shot learning. Methods with $(\cdot)^{\dagger}$ adopted ResNet18 as backbone with input size 224×224, while $(\cdot)^{*}$ used ResNet18 with 84×84. The rest methods used ResNet12 with 84×84 as in (Wang et al. [190]). Results of previous works are from (Wang et al. [190]).

Few-Shot Learning. In addition to transfer clustering, we also compared the proposed SUPREME method on a downstream transfer learning task, Few-shot learning. We compared SUPREME with TapNet (Yoon et al. [210]), CTM (Li et al. [104]), MetaOpNet (Lee et al. [101]), TPN (Liu et al. [109]), TEAM (Qiao et al. [142]) and ICI (Wang et al. [190]) on two challenging FSL benchmarks. As shown in Table 5.2, the performance advantages achieved by the transductive methods over the inductive ones were marginal due to the limited amount of unlabelled target samples (15 samples per class) that were available for the transductive approaches under the meta-learning-based evaluation protocol. Even though SUPREME made use of none human annotation in target domains, it yielded superior performance in most of the cases even when being compared with FSL methods trained with 5 labelled samples per target category (5-shot). The superiority achieved by SUPREME demonstrates that it is possible to effectively transfer knowledge across domains without human annotation and the transferred model should is potential to benefit FSL by further trained with the task-specific assumption, *i.e*. label of target samples.

To study the effect of unlabelled target data amount to our model's discrimination ability, we varied it from 15 (same as in FSL) to 600 (maximal on Mini-Imagenet) and reported the clustering accuracy. As shown in Figure 5.4, the more visual information we can use in model training, the

better our model will be. Besides, even when we use the same amount of unlabelled target data as FSL approaches, our model can still produce competitive results. Although the transfer clustering setting makes use of the whole unlabelled target set to learn the data partitions, considering the high cost of human annotation, SUPREME should not be less applicable than FSL methods in real-world scenarios by requiring no manual label.

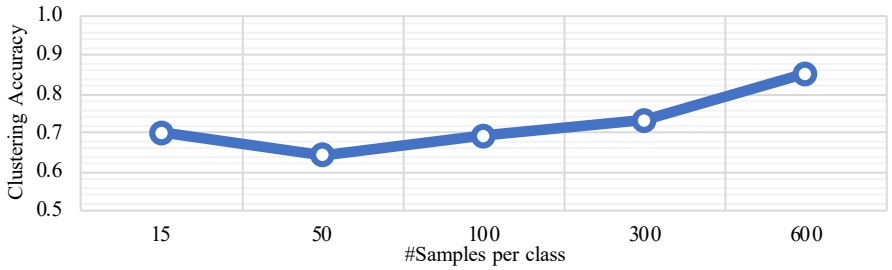


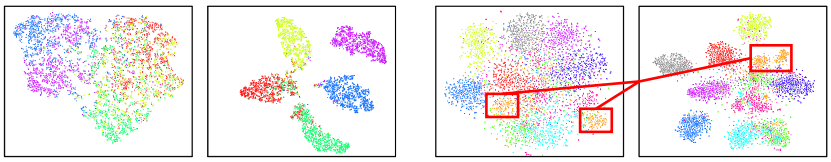
Figure 5.4: Effect of unlabelled target domain size to model discrimination ability. We randomly sampled variant amount {15, 50, 100, 300, 600} of data from each of the 5 target categories and report the clustering accuracy.

Protocol	Method	CUB	FLO	SUN	AWA2
Inductive	ALE	0.549	0.485	0.581	0.599
	CLSWGAN	0.573	0.672	0.608	0.682
	SE-GZSL	0.596	-	0.634	0.692
	Cycle-CLSWGAN	0.586	0.703	0.599	0.668
	F-VAEGAN-D2	0.610	0.677	0.647	0.711
Transductive	ALE-tran	0.545	0.483	0.557	0.707
	GFZSL	0.500	0.854	0.640	0.786
	DSRL	0.487	0.577	0.568	0.728
	UE-finetune	0.721	-	0.583	0.797
	F-VAEGAN-D2	0.711	0.891	0.701	0.898
	SUPREME	0.686	0.690	0.408	0.831

Table 5.3: Comparisons to the state-of-the-art methods in zero-shot learning. Results of previous works are from (Xian et al. [198]).

Zero-Shot Learning. We further compared the proposed SUPREME method in ZSL with ALE (Akata et al. [2]), CLSWGAN (Xian et al. [197]), SE-GZSL (Kumar Verma et al. [95]),

Cycle-CLSWGAN (Felix et al. [47]), F-VAEGAN-D2 (Xian et al. [198]), ALE-tran (Xian et al. [196]), GFZSL (Verma and Rai [183]), DSRL (Ye and Guo [209]) and UE-finetune (Song et al. [164]). Table 5.3 shows that the superiority of the transductive methods over the inductive ones in ZSL was more significant than that in FSL due to the sufficient unlabelled target data available in model learning. The competitive results yielded by SUPREME on all the datasets except SUN, especially against the inductive ZSL setting (more realistic and general), demonstrated its compelling discrimination ability on target domains even without any word vector prototype mapping in the text space or human labelled attribute learning on class description. However, the rather poor performance on SUN, in which the number of target samples is limited to around 1,000 but the target classes are larger than other benchmarks, implies that sufficiently large unlabelled training data is important to our model. The experimental results, once again, demonstrate our model's potential to benefit another downstream transfer learning task.



(a) CIFAR10 (b) CIFAR100

Figure 5.5: Visualisations of target samples' representation before and after transferring by SUPREME. The left in each pair of images was the feature space produced by the model pretrained on source data, while the right images was by SUPREME. Ground-truth labels were colour-coded.

5.4.4 Visualisation and Qualitative Study

We visualised the representations of target samples by t-SNE (Maaten and Hinton [116]) in both the pretrained and transferred feature spaces. Figure 5.5 shows that, even though the target samples can roughly form some groups in the pretrained feature space, the boundaries were ambiguous. As our objective function encourages determined assignments, which means no samples should hold similar probabilities to be assigned into multiple clusters, our models yielded clusters in higher compactness and discrimination. Moreover, as shown in the red boxes highlighted in Figure 5.5 (b), thanking to the auxiliary self-supervision constructed by intrinsic information on target data, our method was able to correctly cluster samples of the same classes but were initially far away.

(a) Positive cases (b) Negative cases (c) Confusion matrices (d) Class-wise accuracy

Figure 5.6: Qualitative studies on CIFAR-10. **(a)** Samples that were wrongly classified by k-means but corrected by SUPREME. **(b)** Samples that were wrongly classified by SUPREME. **(c)** Confusion matrices yielded by k-means (left) or SUPREME (right). **(d)** The clustering accuracy of k-means or SUPREME on each target category.

Cases study. We randomly selected 6 samples that were mistakenly classified into each of the target categories by k-means but assigned into the correct clusters by SUPREME, *e.g.* all the samples from the first row in Figure 5.6 (a) were classified as "Dog" by k-means but assigned into their ground-truth classes respectively by SUPREME. Due to the discrepancy between domains, the pretrained model failed to distinguish samples from different target classes especially when handling classes with large discrepancy to the source ones. By taking the self-supervision into concern, SUPREME learns not only from the prior-knowledge but also the visual similarity that is intrinsically available in the target domain and that enables it to better describe the target distribution. We then selected another 6 instances from each target category at random and all of them were wrongly classified by both k-means and SUPREME. As shown in Figure 5.6 (b), our model failed to make correct predictions in extreme cases like large inter-class similarity (boat with wheels was mistaken as truck), unexpected textures (bees-clothing dog), *etc*.

Confusion matrices. Figure 5.6 (c) shows the confusion matrices between the ground-truth memberships and the assignments produced by k-means (left) or SUPREME (right). It demonstrates that k-means usually confused among dog, frog and horse (red border) or ship and truck (orange border) but rarely mistook the latter group as the former one. This was because the source domain was also consisting of animals and manufactures, so the pretrained model has learned to differentiate between them. However, the class-specific patterns that tell the source categories apart were different from those of the target ones, *i.e.* nontransferable. Therefore, k-means is less sensitive to visual discrepancies between classes in finer grained. To address such a limitation, SUPREME adapts the pre-learned knowledge to the target domain and complements the non-transferable part by self-supervision, which enables it to better identify the intra-class

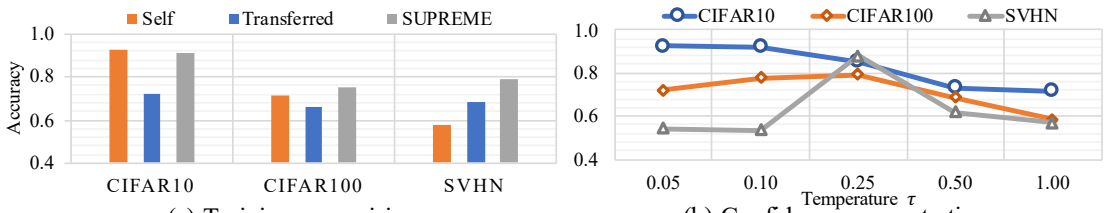
difference and inter-class similarity.

Class-wise performance. We then studied how well the models perform on each target class. As indicated in Figure 5.6 (d), k-means showed inconsistent capacity on different categories and did worst on recognising frog which is likely least visually related to the source categories including airplane, automobile, bird, cat and deer. The SUPREME model not only boosted the performance on the categories that are closely related to the source ones but also did well on those didn't when the pre-learned knowledge are less transferable. This demonstrates the effectiveness of our auxiliary learning design to learn from the self-supervision constructed according to the visual characteristic of target samples and take it as the remedy of the transferred supervision.

5.4.5 Component Analysis and Discussions

We conducted detailed ablation studies to investigate the effectiveness of different design choices in our model for in-depth analysis.

Transferred supervision v.s. Self-supervision. As our key idea holds the assumption that self-supervision can provide complementary constraints to the target samples to which the prior-knowledge is not applicable, we evaluated the effectiveness of both the transferred supervision and self-supervision to better understand their individual contributions to the model. According

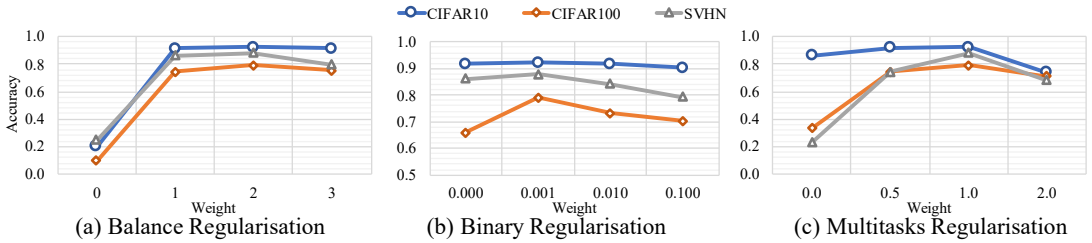


(a) Training supervisions (b) Confidence concentration

Figure 5.7: Ablation studies of training objectives. **(a)** Decouple different supervisions to investigate their individual contribution to model's capacity. **(b)** Effect of temperature τ which decides the concentration degree of confidence distribution.

to Figure 5.7 (a), both the prior-knowledge and the self-supervision are able to provide useful constraints independently in model training. Our SUPREME model achieves better performance in most cases, which indicates these two supervisions can benefit each other. It is also interesting to see that the self-supervised model marginally surpasses SUPREME on CIFAR-10. This suggests that the prior-knowledge from the source domain sometimes contains misleading and inaccurate (non-transferable) information of high confidence to the target domain.

Confidence Distribution Concentration. The temperature τ used to compute the confidence of constraints in Eq. 5.4 decides the concentration degree of the normalised confidence distribution, hence, it can be interpreted as the reliability of prior-knowledge in terms of self-supervision. As shown in Figure 5.7 (b), our model is able to attain promising performance with a wide range of τ but the best results on different datasets can sometimes be sensitive to it, *e.g.* $\tau = 0.25$ in SVHN. This is because measuring the relevance of two domains and the transferability of prior-knowledge is intrinsically challenging. Thereby, the setting of τ is intricately related to various factors, *e.g.* capacity of pretrained models. Due to the existence of human annotations on source domains, a reasonable setting of τ can be determined by cross-validation, which is a conventional solution in transfer learning (Xian et al. [196]).



(a) Balance Regularisation (b) Binary Regularisation (c) Multitasks Regularisation

Figure 5.8: Effects of training regularisations in SUPREME. **(a)** Balance regularisation on clusters size distribution. **(b)** Binary regularisation on common factor space. **(c)** Multitasks regularisation to avoid trivial representation.

Effects of regularisations. The SUPREME model is trained with several regularisations and we investigated their necessity as well as our model's robustness to them by varying their weights within different ranges. As shown in Figure 5.8, the significant performance drop in most cases caused by removing either of these regularisations (setting the weight to 0) demonstrate that none of them is redundant for effective knowledge transfer. Furthermore, the stability and the similar trends in parameter values on different datasets indicate the scalability and robustness of our model which requires no exhaustive parameter tuning.

5.5 Summary

In this chapter, we addressed a common underlying problem among several unsupervised transfer learning tasks that all aiming to model a discriminative latent space in an unsupervised (or weak semi-supervised) manner with the help of the prior-knowledge acquired from labelled data in related domains. To that end, we propose a new generic and more scalable solution to unsupervised

transfer learning by formulating a *self-SUPervised REMEdy* method to augment transfer clustering with self-supervised learning. It is always the case that some of the target samples will fail to yield reliable transferred supervision from prior-knowledge due to distribution shift/discrepancy, the proposed SUPREME method is designed to identify those "hard" target samples and provide them with self-supervision based on intrinsic pairwise similarities among the target images in relation to the source domain pre-knowledge. Experiments on extensive transfer clustering, FSL and ZSL benchmark datasets demonstrate the compelling performance of the proposed method against a wide range of the state-of-the-art models in all three tasks. Uniquely, SUPREME shows competitive discrimination ability on both ZSL and FSL tasks *without* utilising any additional assumptions on sample-wise human annotation or class-level semantic knowledge, as the case in all conventional ZSL and FSL methods. Detailed ablation studies and in-depth analysis are provided to give insights on SUPREME design considerations.

Chapter 6

Propagating Knowledge from Incomplete Labels in Video Activity Localisation

---

This chapter studies knowledge propagation from incomplete labels in video activity localisation by natural language when the exhaustive temporal boundary of the moments-of-interest (MoIs) are collective (*i.e.*, the natural language queries are associated with videos rather than video segments) or when they are uncertain with subjective annotation bias. Given an untrimmed and unstructured video V composed of L^f frames $V = \{\boldsymbol{x}_i^f\}_{i=1}^{L^f}$ and a natural language sentence $Q = \{\boldsymbol{x}_i^w\}_{i=1}^{L^w}$ with L^w words describing the activity depicted in a specific segment of the video V, the objective of video activity localisation is to learn discriminative representations of video frames/segments and align them with the query sentences so that the model can predict the temporal boundary of a target moment (S,E) to ensure that the video segment $\{\boldsymbol{x}_i^f\}_{i=S}^{E}$ is matching with the query Q in semantics. Labelling video activities is fundamentally a more challenging task than conventional image- and video-level annotations, which needs the annotators to go through the long and untrimmed videos frame-by-frame to identify the proper boundaries that are inherently ambiguous. In this chapter, we present a novel *Cross-sentence Relations Mining* method to learn the matching relationships of video segments and sentence descriptions when only the labels on video-sentences association are available but not the temporal boundary annotations on video segments. Moreover, a new *Elastic Moment Bounding* approach is introduced to deal with the inherent uncertainties in boundary labels, which helps derive universal visual-textual correlations and locate video activity more accurately in inference.

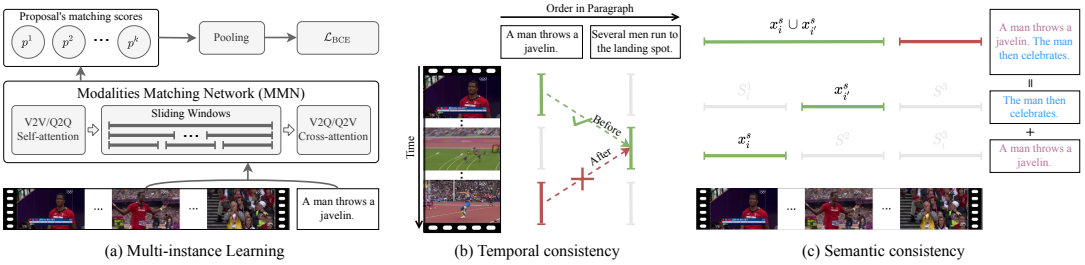
This chapter is organised as follows: Section 6.1 discuss weakly-supervised video activity localisation when the video-text labels are provided in collective, where Section 6.1.1 gives problem statement and approach overview, Section 6.1.2 introduces the conventional multi-instance learning based design for locating activity without boundary labels and Section 6.1.3 elaborates the proposed mechanism of cross-sentence mining while model training and optimisation are given in Section 6.1.4; Section 6.2 provides extensive experiments for evaluating the CRM model with detailed ablation studies and in-depth analysis. Section 6.3 investigates the uncertainty in temporal annotations when being used as supervisions in model learning, with the problem statement and approach overview given in Section 6.3.1. Section 6.3.2 and Section 6.3.3 in respectively present a general framework for frame-wise temporal boundary prediction and a new strategy to improve model's robustness to labelling uncertainty by simultaneously learning video-text matching relationships in segment-level. Section 6.3.4 introduce how the EMB model should be trained and used in inference; Section 6.4 evaluates EMB thoroughly and investigate the effectiveness of each of its components; Section 6.5 presents a summary of this chapter.

6.1 Knowledge Propagation from Collective Labels

This section presents the proposed *Cross-sentence Relations Mining* model for weakly-supervised video activity localisation to study knowledge propagation from collective labels.

6.1.1 Problem Statement and Approach Overview

Problem statement. Suppose for each video, we have a description paragraph consisting of L^q text query sentences $\mathcal{Q} = \{Q^j\}_{j=1}^{L^q}$ one-to-one describing the MoIs in V. Given a video-query pair (V, Q^j), by dividing the untrimmed video V into L^s candidate segments $\{\boldsymbol{x}_i^s\}_{i=1}^{L^s}$ using sliding windows (Lin et al. [107]; Ma et al. [115]) as the *proposals*, our objective in weakly-supervised activity localisation is to select the $\boldsymbol{x}_i^s$ which is most aligned with Q^j *in semantic*. Although the video-paragraph (multi-sentences) relations are available in training, there is no access to the ground-truth per-sentence temporal boundary, *i.e.*, the manual labels are given in collective. This is a weakly-supervised learning problem where video proposals $\boldsymbol{x}_i^s$ interact with the text queries Q^j to discover the most plausible matches between video segments and text sentences while the labels are given in video-level on associating Q^j with V rather than $\boldsymbol{x}_i^s$.



(a) Multi-instance Learning (b) Temporal consistency (c) Semantic consistency

Figure 6.1: An overview of the proposed *Cross-sentence Relations Mining* method.

Approach Overview. In this section, we formulate a *Cross-sentence Relations Mining* method for this task. Figure 6.1 shows an overview. We first learn the visual-text alignment in video-level with the same spirit of Multi-Instance Learning to feed a video-query pair into a *Modalities Matching Network* (MMN), which predicts the matching score of the query and every proposal and supervise the max-pooling of scores by binary cross-entropy loss (Figure 6.1 (a)). We then explore the order of two descriptions in the paragraph and optimise their joint matching scores to a proposals pair with consistent temporal relations (Figure 6.1 (b)). Furthermore, we synthesise a longer query by forming pairs of sentences in a paragraph (concatenation) and encourage its pairwise localisation to be semantically consistent with the union of proposals individually selected for each sentence (Figure 6.1.1 (c)). This is to minimise the ambiguities in sentences so to improve the model's interpretation of multiple video moments in a more complex sentencing context.

6.1.2 Video-Sentence Alignment

We start with the alignment of representations from two different modalities, *i.e.*, an untrimmed video $V = \{\boldsymbol{x}_i^f\}_{i=1}^{L^f} \in \mathbb{R}^{D^v \times L^f}$ composed of L^f frames and a query sentence $Q^j = \{\boldsymbol{x}_{j,i}^w\}_{i=1}^{L^w} \in \mathbb{R}^{D^t \times L^w}$ with L^w words. To explore the relation of V and Q^j and enable video-text interaction, both the video and sentence are first projected into D-dimensional spaces by two independent fully-connected layers, respectively. For clarity concern, we reuse the symbols $V \in \mathbb{R}^{D \times L^v}$ and $Q^j \in \mathbb{R}^{D \times L^w}$ after projections. The video V and the query Q^j will then be fed into a MMN which will generate a set of candidate moments (proposals) $\{\boldsymbol{x}_1^s, \boldsymbol{x}_2^s, \cdots, \boldsymbol{x}_{L^s}^s\}$ by sliding windows (Lin et al. [107]; Ma et al. [115]) and predicts their individual matching scores with the input query $\{p(\boldsymbol{x}_i^s|Q^j)\}_{i=1}^{L^s}$ (Figure 6.1 (a)). Motivated by the remarkable success of Transformer (Vaswani et al. [182]; Devlin et al. [38]) on sequence analysis, the MMN is composed of a stack of attention

units to explore both the within and cross-modal correlation.

Attention Unit. As the building block of our MMN, the attention unit plays a significant role to learn the representation of a target sequence in terms of its correlations with every element in a reference sequence. For instances, the target and reference sequences can be two videos (V_i, V_j), two sentences (Q_i, Q_j) or a video-sentence pair (V_i, Q_j). Given a target sequence $X^t \in \mathbb{R}^{D\times L^t}$ and a reference $X^r \in \mathbb{R}^{D\times L^r}$, an attention unit $\delta(X^t, X^r)$ attends X^t using X^r as follows:

$$\begin{aligned} \mathcal{A} &= {X^t}^\top W^q W^k {X^r}^\top / \sqrt{D} \in \mathbb{R}^{L^t \times L^r} \\ \delta(X^t, X^r) &= \mathrm{FC}\big(X^t + W^v X^r \mathrm{Softmax}(\mathcal{A})^\top\big) \in \mathbb{R}^{D\times L^t}. \end{aligned} \tag{6.1}$$

The notions $\{W^q, W^k, W^v\} \in \mathbb{R}^{D\times D\times 3}$ in Eq. (6.1) are three learnable matrices and the coefficient $1/\sqrt{D}$ is to counteract the effect of small gradients caused by large D (Vaswani et al. [182]). The $\mathrm{Softmax}(\cdot)$ is the row-wise softmax normalisation and $\mathcal{A}$ is the correlation scores of target-reference element pairs. The $\mathrm{FC}(\cdot)$ function is a linear projection with consistent input-output dimensions. The attended result serves as the updated representation of the target sequence.

To investigate the video-text matching relations, it is essential to explore not only the within-modal context but also the cross-modal interaction (Ma et al. [115]). Hence, the MMN is constructed by both self-attention and cross-attention blocks. The video V and the query Q^j are first fed into two independent self-attention blocks respectively, in which the target and reference inputs are from the same modalities:

$$V \leftarrow \delta^{\mathrm{V2V}}(V, V), \quad Q^j \leftarrow \delta^{\mathrm{Q2Q}}(Q^j, Q^j). \tag{6.2}$$

By doing so, the salient clips/words in the input video/query are highlighted by considering the context of the video or sentence. Conventional sliding window strategy (Lin et al. [107]; Ma et al. [115]) is then adopted to divide the video into L^s proposals $V = \{\boldsymbol{x}_i^s\}_{i=1}^{L^s} \in \mathbb{R}^{D\times L^s}$. Each proposal is composed of arbitrary continual frames in $\{\boldsymbol{x}_i^f\}_{i=1}^{L^f}$ and represented by max-pooling their features. After that, the two representations are interacted by cross-attention:

$$V \leftarrow \delta^{\mathrm{Q2V}}(V, Q^j), \quad Q^j \leftarrow \delta^{\mathrm{V2Q}}(Q^j, V), \tag{6.3}$$

which attends one modality by another so to suppress the redundant text and irrelevant visual information in query sentence and target video, respectively.

Matching Score. Given the video segment features $V = \{\boldsymbol{x}_i^s\}_{i=1}^{L^s}$ and the text representation $Q^j = \{\boldsymbol{x}_{j,i}^w\}_{i=1}^{L^w}$, the matching score $p(\boldsymbol{x}_i^s | Q^j)$ of a proposal-query pair is predicted according to

both the modalities. The sentence representation is first computed by aggregating all the words: $\boldsymbol{x}_j^q \leftarrow \text{cmax}(Q^j) \in \mathbb{R}^{D\times 1}$ where $\text{cmax}(\cdot)$ denotes the column-wise max-pooling function, which is then fused with every proposal's representation (Gao et al. [50]; Hendricks et al. [71]):

$$\boldsymbol{x}_{i,j} = (\boldsymbol{x}_i^s + \boldsymbol{x}_j^q) \| (\boldsymbol{x}_i^s \otimes \boldsymbol{x}_j^q) \| \text{FC}(\boldsymbol{x}_i^s \| \boldsymbol{x}_j^q). \tag{6.4}$$

The notion $(\cdot \otimes \cdot)$ indicates the element-wise multiplication and $(\cdot \| \cdot)$ is the concatenation of two vectors while $\text{FC}(\cdot)$ standing for a linear projection. After that, the joint representations $\{\boldsymbol{x}_{i,j}\}_{i=1}^{L^s}$ are fed into a linear classifier:

$$p(\boldsymbol{x}_i^s | Q^j) = \sigma(\boldsymbol{x}_{i,j} W^\top + B). \tag{6.5}$$

The variable $\{W, B\}^\top \in \mathbb{R}^{D+1}$ is the weights of classifier and $\sigma(\cdot)$ is the sigmoid function. The yielded probabilities $\{p(\boldsymbol{x}_i^s | Q^j)\}_{i=1}^{L^s} \in (0,1)$ serve as the matching scores between proposals and query, which is abbreviated to $p_{i,j}$. In inference, the most aligned proposal can be obtained in a maximum likelihood manner:

$$\boldsymbol{x}_*^s = \arg\max_{\boldsymbol{x}_i^s} p_{i,j}, \tag{6.6}$$

whose boundary $(\tilde{S}, \tilde{E})$ is predicted as the temporal boundary of the activity corresponding to Q^j.

Multi-Instance Learning. In the absence of temporal boundary, the ground-truth moment is agnostic. Therefore, we optimise the matching scores in video-level to facilitate video-text alignment. To that end, the matching score between the video V and the query Q^j is obtained by the max-pooling of all the proposals' score $p(V|Q^j) \leftarrow \max(\{p_{i,j}\}_{i=1}^{L^s})$. For each positive pair (V, Q^j) given manually on the dataset, we construct two negative counterparts by replacing either V or Q^j by a randomly sampled video V^- or sentence Q^- from the mini-batch and compute their matching scores in the same way as $p(V|Q)$. The BCE loss function is then adopted as the video-query alignment supervision signal:

$$\mathcal{L}_{\text{BCE}}(V, Q^j) = -2\log\big(p(V|Q^j)\big) - \log\big(1 - p(V|Q^-)\big) - \log\big(1 - p(V^-|Q^j)\big), \tag{6.7}$$

where the coefficient 2 is applied to the positive term considering the balance of positive and negative pairs. The rationale behind Eq. (6.7) is assuming that the MoIs in one video doesn't exist in any other videos so (V, Q^-) and (V^-, Q^j) should be semantically unmatched. By minimising $p(V|Q^-)$ and $p(V^-|Q)$, the predictions of the incorrect proposals in V with different semantics from Q^j will also be minimised implicitly so that the learned matching scores can reveal the inherent visual-textual correlation. This takes the spirit of MIL (Keeler et al. [88]) by treating the proposals as the instances in a bag (video) and learning with the bag-level annotations.

6.1.3 Cross-Sentence Relations Mining

The $\mathcal{L}_{\text{BCE}}$ in Eq. (6.7) aligns queries with the proposals yielding the largest matching scores among all the candidates. However, the predicted scores can be unreliable due to the visually indiscriminate moment proposals existed in videos and text ambiguities in individual sentences which will lead to video-text misalignment in training. Therefore, we explore the cross-sentence relations to select reliable proposals with consistent cross-moment relations.

Temporal Consistency. As the video frames are naturally exhibited to the viewers in time order, the temporal relations of different MoIs should intrinsically be encoded in the order of their descriptions in the paragraph. With such an assumption, we can identify the pairs of proposals both yielding high predicted matching score with the corresponding queries but inconsistent in temporal relations, which are likely to be incorrect. Given arbitrary query sentences pair $(Q^j, Q^{j'})$ from the description paragraph of video V, their respectively selected segments $(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s)$ should satisfy similar temporal structure with them, *i.e.*, $\boldsymbol{x}_i^s$ should occur before $\boldsymbol{x}_{i'}^s$ in the video if Q^j is in front of $Q^{j'}$ in the paragraph and vice versa. The temporal order of two proposals $\mathcal{R}(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s) = 0$ if $\boldsymbol{x}_i^s$ starts before $\boldsymbol{x}_{i'}^s$ in the video, otherwise $\mathcal{R}(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s) = 1$. Similarly, $\mathcal{R}(Q^j, Q^{j'}) = \mathbb{1}[j \geq j']$ where j and j' are the position of sentences in the paragraph. The temporal (TMP) constraint is then formulated to ensure $\mathcal{R}(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s) = \mathcal{R}(Q^j, Q^{j'})$.

By assuming the matching scores of different queries to any proposals are independent, the joint probability of Q^j and $Q^{j'}$ are respectively matching with $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_{i'}^s$ is:

$$p(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s | Q^j, Q^{j'}) = p(\boldsymbol{x}_i^s | Q^j) \cdot p(\boldsymbol{x}_{i'}^s | Q^{j'}). \tag{6.8}$$

As shown in Figure 6.1 (b), we take the queries' order as the ground-truth for the temporal relation of the proposal pair. Given Q^j and $Q^{j'}$, the joint probabilities set $\{p(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s | Q^j, Q^{j'})\}_{i,i'=1}^{L^s}$ is then divided into two subsets: for all the proposal pairs $(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s)$, the joint probability $p(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s | Q^j, Q^{j'}) \in P_t^+$ if $\mathcal{R}(\boldsymbol{x}_i^s, \boldsymbol{x}_{i'}^s) = \mathcal{R}(Q^j, Q^{j'})$, otherwise belonging to P_t^-. The MIL loss is re-formulated with the temporal constraint:

$$\mathcal{L}_{\text{TMP}}(V, Q^j, Q^{j'}) = -\log\big(\max(P_t^+)\big) - \log\big(1 - \max(P_t^-)\big). \tag{6.9}$$

By training with $\mathcal{L}_{\text{TMP}}$, the model learns to align the proposals with queries only if they are temporally consistent. This refrains the model from visual-textual misalignment in the absence of ground-truth temporal annotations.

Semantic Consistency. To minimise the negative impact from ambiguous per-sentence expressions in isolation and to explore the context of a paragraph, it is beneficial for a model to consider broader semantics beyond individual sentences by relating other expressed objects/actions in a wider context (Mun et al. [123]). However, it is non-trivial to explicitly do so since the object/action's information is missing without fine-grained annotation. In this case, we propose to form pairs of MoIs by concatenation in the same videos: $Q^{j,j'} = Q^j \| Q^{j'}$ and train the model to locate the concatenated longer query with the consideration of both sentences in each pair. Given the proposals $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_{i'}^s$ with the largest $p(\boldsymbol{x}_i^s, \boldsymbol{x}_{k'}^s | Q^j, Q^{j'})$ in Eq. (6.9), the matching scores of $Q^{j,j'}$ and the video segments $\boldsymbol{x}_k^s$ is optimised to encourage the consistency of $\boldsymbol{x}_k^s$ and $\boldsymbol{x}_i^s \cup \boldsymbol{x}_{i'}^s$ (Figure 6.1 (c)). As in the temporal constraint, we divide the predicted scores $p(\boldsymbol{x}_k^s | Q^{j,j'})$ into two subsets: for all the proposals $\boldsymbol{x}_k^s$ in the video V, $p(\boldsymbol{x}_k^s | Q^{j,j'}) \in P_s^-$ if $\text{IoU}(\boldsymbol{x}_k^s, \boldsymbol{x}_i^s \cup \boldsymbol{x}_{i'}^s) < \tau$, and P_s^+ is composed of the $\boldsymbol{x}_k^s$ which is most consistent with $\boldsymbol{x}_i^s \cup \boldsymbol{x}_{i'}^s$. The τ decides how two proposals are deemed consistent regarding their Intersection over Union (IoU) which is set to 0.5 in practice. The constraint on the semantic (SMT) consistency of $\boldsymbol{x}_k^s$ and $\boldsymbol{x}_i^s \cup \boldsymbol{x}_{i'}^s$ is formulated as:

$$\mathcal{L}_{\text{SMT}}(V, Q^j, Q^{j'}) = -\log\big(\max(P_s^+)\big) - \log\big(1 - \max(P_s^-)\big). \tag{6.10}$$

To minimise $\mathcal{L}_{\text{SMT}}$, the model is explicitly trained to consider the semantics of both Q^j and $Q^{j'}$ when locating $Q^{j,j'}$ so to ensure the overlap of $\boldsymbol{x}_k^s$ and $\boldsymbol{x}_i^s \cup \boldsymbol{x}_{i'}^s$. By introducing additional longer queries synthesised from pairwise sentences in model training, it enhances the model's capacity to interpret and match more complex descriptions to video moments, critical in practice due to that untrimmed raw videos are often unstructured.

6.1.4 Model Training

In each training iteration, we randomly sample n_{bs} videos with a pair of queries (Q_i^1, Q_i^2) for each from its paragraph description as a mini-batch and the overall loss to be minimised is:

$$\mathcal{L} = \frac{1}{2 * n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{2} \mathcal{L}_{\text{BCE}}(V_i, Q_i^j) + \frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \mathcal{L}_{\text{TMP}}(V_i, Q_i^1, Q_i^2) + \frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \mathcal{L}_{\text{SMT}}(V_i, Q_i^1, Q_i^2). \tag{6.11}$$

Since the objective function $\mathcal{L}$ in Eq. (6.11) is differentiable, conventional stochastic gradient descent algorithm is adopted for end-to-end model training. The overall process of a training iteration is summarised in Algorithm 5.

Algorithm 5: CRM for video grounding with collective labels.

Input: Untrimmed videos $\mathcal{V}$, Paragraph descriptions $\mathcal{Q}$.

Output: A video activity localisation model.

Sampling a random mini-batch of videos;

Sampling two queries for each video from its paragraph;

foreach video-query pair **do**

  Mapping video and query to D-dimensional spaces by linear projections;

  Conducting V2V and Q2Q self-attention (Eq. (6.2));

  Generating proposals by sliding windows;

  Conducting V2Q and Q2V cross-attention (Eq. (6.3));

  Fusing each proposal's feature with the query (Eq. (6.4));

  Computing the proposal-query matching scores (Eq. (6.5));

  Computing the BCE-based MIL loss $\mathcal{L}_{\text{BCE}}$ (Eq. (6.7));

end

Computing the temporal consistency constraint $\mathcal{L}_{\text{TMP}}$ (Eq. (6.9));

Computing the semantic consistency constraint $\mathcal{L}_{\text{SMT}}$ (Eq. (6.10));

Updating model weights by back-propagation;

6.2 Experiments on Knowledge Propagation from Collective Labels

6.2.1 Datasets and Metrics

Datasets. Experiments were conducted on two video activity localisation datasets: (1) Charades-STA (Gao et al. [50]) contains 12,408/3720 video-query pairs from 5338/1334 videos for training and testing, respectively. The query sentences are composed of 7.2 words on average and the average duration of the target video moments and untrimmed videos are 8.2 and 30.6 seconds; (2) ActivityNet-Captions (Krishna et al. [91]) is a much larger-scale dataset composed of 19,290 videos with 37,421/17,505/17,031 MoIs in the train/val_1/val_2 split. The average length of queries is 14.8 words while that of the MoIs and untrimmed videos are 36.2 and 117.6 seconds.

The activities captured in those two datasets are of various complexity: Only 6% of the descriptions involve more than one actions in Charades whilst 44% in ActivityNet with 12% *vs.* 44% regarding the number of people (Lei et al. [103]).

Evaluation metrics. We followed previous works (Duan et al. [43]; Wu et al. [194]; Chen and Jiang [26]) to evaluate the activity localisation results by the "IoU@m" metric where m is the pre-defined temporal IoU thresholds. Given the temporal boundary (S,E) of a target moment and the selected segment proposal bounding at $(\tilde{S},\tilde{E})$ (Eq. (6.6)) with the largest predicted matching score, the IoU between the two video segments is computed by $\frac{\max(0,\min(E,\tilde{E})-\max(S,\tilde{S}))}{\max(E,\tilde{E})-\min(S,\tilde{S})}$. A prediction is considered correct if its IoU with the ground-truth is greater than the pre-defined IoU thresholds m set to $\{0.1,0.3,0.5\}$ on ActivityNet and $\{0.3,0.5,0.7\}$ on Charades (Duan et al. [43]; Wu et al. [194]).

6.2.2 Implementation Details

We used VGG (Simonyan and Zisserman [163]) (4096-D) and ResNet152 (He et al. [68]) (2048-D) feature representations officially released with the datasets for per-frame representations in Charades and ActivityNet, respectively. The videos were truncated evenly (and zero-padded) into 128 clips in Charades and 256 in ActivityNet, with each clip represented by the max-pooling of 5 continual frame's features as the minimal video unit ($\{\boldsymbol{x}_i^f\}_{i=1}^{L^f}$). The pre-trained GloVe embedding (Pennington et al. [140]) was adopted as the word feature representation (300-D) and the maximal sentence length was set to 20 words. Both the clip and word representations were linearly mapped to 256-D spaces before being fed into MMN. The sliding windows stride was 8 and the window sizes were $\{8,12,20,32,64\}$ in Charades and $\{8,16,32,64,128\}$ in ActivityNet. The temporal dependencies of video segments in terms of the same query sentences were explored by an additional self-attention unit before predicting their matching scores. As the paragraph descriptions were pre-divided into individual sentences on both datasets, we restored the order of sentences in the paragraph by the ground-truth start time of MoIs. Note that timestamps were unavailable in proposal selections, neither in training nor testing. The proposed CRM was trained 50 epochs by Adam optimiser with a batch size of 64 and learning rate of $1e-4$. Cross-sentence relations were only used in training with no extra computational cost in testing.

6.2.3 Comparisons with the State-of-the-Art

Table 6.1 compares the performance of CRM against the state-of-the-art video activity localisation models including both fully-supervised (DPIN (Wang et al. [185]), 2D-TAN (Zhang et al. [221]), DRN (Zeng et al. [214]), LGI (Mun et al. [123]), HVTG (Chen and Jiang [26])) and weakly-supervised methods (WS-DEC (Duan et al. [43]), WSLLN (Gao et al. [51]), BAR (Wu

Method	Split	Moment	Query	IoU@0.1	IoU@0.3	IoU@0.5
DPIN	val_2	✓	✗	-	62.40	47.27
2D-TAN		✓	✗	-	59.45	44.51
DRN		✗	✗	-	-	45.45
LGI		✓	✗	-	58.52	41.51
HVTG		✗	✗	-	57.60	40.15
WS-DEC	val_1	✗	✗	62.71	41.98	23.34
WSLLN		✗	✗	75.40	42.80	22.70
BAR		✓	✗	-	49.03	30.73
CRM (Ours)		✓	✓	**76.66**	**51.17**	**31.67**
SCN	val_1	✓	✗	71.48	47.23	29.22
RTBPN		✓	✗	73.73	49.77	29.63
CCL		✓	✗	-	50.12	31.07
CRM (Ours)		✓	✓	**81.61**	**55.26**	**32.19**
WS-DEC	OOD	✓	✗	30.17	17.00	7.17
CRM (Ours)		✓	✓	**38.35**	**22.77**	**10.31**

Method	Moment	Query	IoU@0.3	IoU@0.5	IoU@0.7
DPIN	✓	✗	-	47.98	26.98
2D-TAN	✓	✗	-	39.81	23.25
DRN	✗	✗	-	53.09	31.75
LGI	✓	✗	72.96	59.46	35.48
HVTG	✗	✗	61.37	47.27	23.30
TGA	✗	✗	29.68	17.04	6.93
SCN	✓	✗	42.96	23.58	9.97
LoGAN	✓	✗	51.67	34.68	14.54
BAR	✓	✗	44.97	27.04	12.23
RTBPN	✓	✗	**60.04**	32.36	13.24
VLANet	✓	✗	45.24	31.83	14.17
CCL	✓	✗	-	33.21	15.68
CRM (Ours)	✓	✓	53.66	**34.76**	**16.37**

Table 6.1: Performance comparisons of CRM on ActivityNet-Captions **(left)** and Charades-STA **(right)**. Fully and weakly-supervised methods are shown in the upper and lower part of each table, respectively. The 'Moment' column refers to methods trained by exploiting multiple video moments corresponding to the same-sentence, whilst the 'Query' column refers to training by cross-sentence temporal ordering and sentence pairing in the context of a paragraph. The 'Split' column denotes the different data splits in the ActivityNet-Captions used in the evaluations. The discounted recall rates (Yuan et al. [213]) are reported for the 'OOD' split of ActivityNet-Captions.

et al. [194]), SCN (Lin et al. [107]), RTBPN (Zhang et al. [223]), CCL (Zhang et al. [224]), TGA (Mithun et al. [121]), LoGAN (Tan et al. [173]), VLANet (Ma et al. [115])). We have the following observations:

- Not surprisingly, fully-supervised models outperform weakly-supervised models clearly by benefitting from the exhaustive temporal annotation on activities' boundary. However, CRM reduces that performance gap by over 41% on the ActivityNet (IoU@0.3).

- Discovering different video moments correlating to the *same-sentence* for proposal selection has been exploited to a good effect by existing methods in an implicit form of attention (Lin et al. [107]; Ma et al. [115]) or 2D temporal convolution (Zhang et al. [223, 221]). However, the notably better performance of CRM compared to those methods further demonstrates the additional advantage of using *cross-sentence* temporal and semantic relations within a paragraph to explicitly constrain *cross-moment* relations for learning better video-text alignment and benefiting per-sentence localisation in testing.

- CRM surpasses the state-of-the-art weakly-supervised methods across the board except for IoU@0.3 on Charades. This demonstrates compellingly the effectiveness of CRM from modelling explicitly cross-sentence relations. Our advantages on the Out-of-Distribution (OOD) split of ActivityNet-Captions (Yuan et al. [213]) further indicate CRM's better multi-modal understanding rather than driven by annotation biases.

6.2.4 Component Analysis

We investigated the effects of different components in CRM model design to study their individual contributions. The "val_1" split of ActivityNet was adopted.

Effects of cross-sentence relations. We evaluated the effectiveness of imposing cross-sentence relational consistency by training the baseline model (BCE) with either the temporal (BCE+TMP) or semantic (BCE+SMT) constraint as well as with both (BCE+TMP+SMT). Figure 6.2 shows that both constraints are beneficial individually and the benefits become more clear when they are jointly adopted. Moreover, the performance improvement is more significant on ActivityNet than Charades. Given the generally more complex activities in ActivityNet, this shows that training CRM on combinations of pairwise sentencing as semantic consistency constraint (Eq. (6.10)) has its unique advantages in activity localisation against more complex query descriptions.

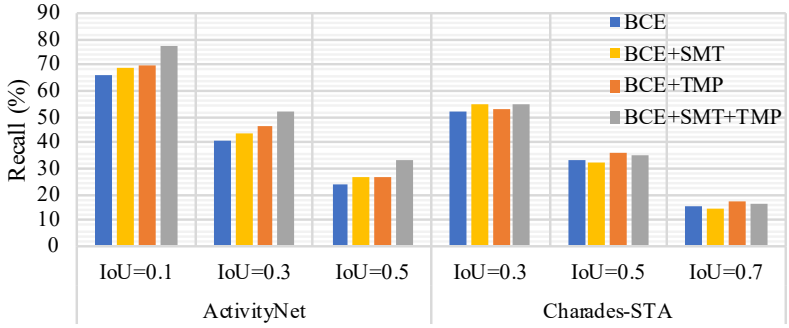


Figure 6.2: Effect of cross-sentence relations mining in CRM. BCE is the base model trains with only the MIL objective (Eq. (6.7)). TMP and SMT are the proposed constraints on temporal (Eq. (6.9)) and semantic (Eq. (6.10)) relational consistency.

Temporal consistency. To verify our assumption on temporal order, we compared how many correct predictions learned with and without $\mathcal{L}_{\text{TMP}}$ (Eq. (6.9)) against the ground-truth. Specifically, for each video consists of n MoIs, we constructed C_n^2 MoI pairs and measured the ratio of consistent pairs by comparing the order of the two ground-truth moments and that of the selected proposals. Table 6.2 shows that by explicitly training CRM with cross-sentence temporal order

Temporal	Train		Test	
	ActivityNet	Charades	ActivityNet	Charades
✗	64.28	73.88	45.02	73.91
✓	82.43	74.88	70.82	74.65

Table 6.2: Temporal consistency between the descriptions of MoI pairs and their selected proposals. Metric: accuracy.

constraint, the video segments selected by CRM is much more consistent in temporal relations on ActivityNet than the base models without it. Although different moments in the test set are localised independently, such advantages are still clear. Besides, it is surprising to see that the cross-moment temporal relations yielded by the base model on Charades are reasonably consistent with the true order but the temporal constraint still benefited the localisation results. This implies the potential advantages of optimising joint matching scores of moment pairs with their descriptions in learning effective visual-textual correlation.

Semantic consistency. As in the analysis of temporal consistency, we enumerated all the possible MoI pairs in the same videos and quantify the semantic consistency by taking the union of MoI pairs as the ground-truth moment corresponding to the concatenation of their descriptions. More specifically, given the sentence description of two MoIs and their selected segments $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_j^s$, we concatenated the two per-sentence queries and identified the video segment $\boldsymbol{x}_k^s$ yielding the largest matching scores with the concatenation. We then computed the temporal IoU between $\boldsymbol{x}_i^s \cup \boldsymbol{x}_j^s$ and $\boldsymbol{x}_k^s$, where $\boldsymbol{x}_k^s$ is deemed semantically consistent with $\boldsymbol{x}_i^s \cup \boldsymbol{x}_j^s$ if $\text{IoU}(\boldsymbol{x}_i^s \cup \boldsymbol{x}_j^s, \boldsymbol{x}_k^s) > 0.5$. Note that it is not necessary for the two moments to be consecutive in time so that our semantic assumption can hold, as the boundary defined by the concatenated description always matches their temporal union. Table 6.3 shows that the baseline model trained without semantic constraint in Eq. (6.10) yields sensible performances in locating the paired queries. This demonstrates that CRM implicitly learns to consider the semantic context of queries by the attention units. The superior results of CRM trained with explicit semantic constraint shows that it encourages broader consensus in semantics across sentences. This explains why the performance advantages of CRM is more significant when locating more complex activities in ActivityNet.

Effects of attention units. As the building block of our MMN backbone introduced in Section 6.1.2, the attention units play a significant role in exploring the videos and sentences data as

Semantic	Train		Test	
	ActivityNet	Charades	ActivityNet	Charades
✗	55.76	35.34	57.84	31.01
✓	68.14	55.46	71.30	51.33

Table 6.3: Semantic consistency between the union of two MoIs' segments and the one selected for the concatenation of their descriptions. Metric: prediction recall at IoU = 0.5.

well as their correlations. We investigated its effect by comparing the prediction recall of CRM constructed with different numbers of attention units, showing its benefits in sequence analysis and video-text interactions (Figure 6.3). On the other hand, due to the limited video data available for training (10K/5K on ActivityNet/Charades), stacking up attention layers fails to further benefit CRM, leading to model performance degradation possibly due to overfitting.

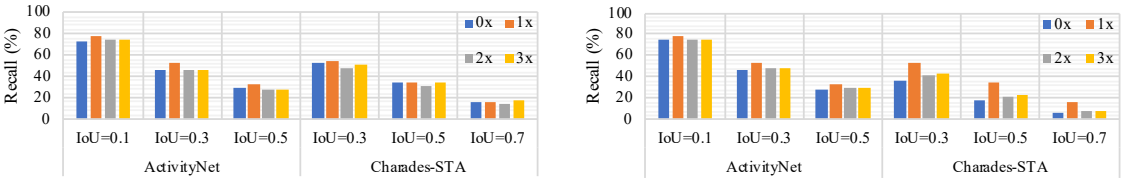


Figure 6.3: Effect of self-attention **(left)** and cross-attention units (right) in CRM. Models are constructed and trained with different numbers of self-attention and cross-attention units to investigate their effects.

Qualitative examples. Figure 6.4 shows some qualitative examples from both ActivityNet and Charades. They show how different MoIs in the same videos may interact with each other so that their relations can be used to optimise per-sentence activity localisation in the context of a paragraph. It is evident that locating video moments by per-sentence independently is unreliable, *e.g.* in the first example (top-row), the man reaches the monkey bars both before and after he walks toward the lower pole. "The first cat" example in the middle-row is ambiguous without context. By explicitly exploring the cross-sentence relations, CRM avoids such ambiguities and minimises video-text misalignment.

6.3 Knowledge Propagation from Uncertain Labels

This section introduces the proposed *Elastic Moment Bounding* model to alleviate the negative impacts of labelling uncertainties in temporal boundary that intrinsically exist in constructing

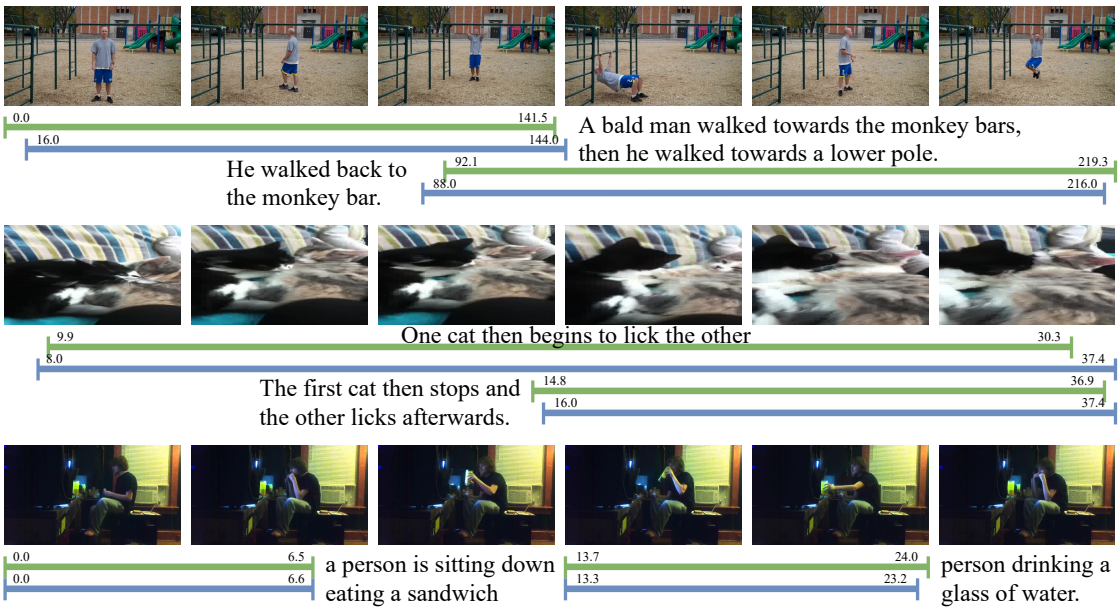


Figure 6.4: Qualitative examples show the interaction between MoIs in the same videos. The green bars indicate the ground-truth MoI's boundaries whilst the blue bars show the model predictions by CRM. The query sentences are simplified for illustrations only.

supervision signals for learning video activity localisation models.

6.3.1 Problem Statement and Approach Overview

Problem statement. Suppose we have a set of long and untrimmed videos $\mathcal{V} = \{V_i\}_{i=1}^{N^v}$ with each composed of L^f frames $V_i = \{\boldsymbol{x}_{i,j}^f\}_{j=1}^{L^f}$. With a set of natural language sentences $\mathcal{Q} = \{Q_i\}_{i=1}^{N^q}$ describing the MoIs in the videos, the objective of fully-supervised video activity localisation is to derive visual-textual correlations from the video-text-boundary triplets $\{\big(Q_i, V_j, (S_j, E_j)\big) \mid i \in [1, N^q] \text{ and } j \in [1, N^v]\}$ so that the model is able to predict accurately the start and end boundary of a moment $(\tilde{S}, \tilde{E})$ in an unseen video V according to a new query sentence Q. Note that different from Section 6.1.1, the activity's descriptions $\mathcal{Q}$ are not organised into paragraphs here so the subscript indicates a single text sentence rather than a paragraph. When studying about the uncertainties in temporal boundary, we assumed the exhaustive temporal labels are available for model training. Considering that semantically similar query sentences are associated with video boundaries with inconsistent visual cues, it is fundamentally challenging to derive universally interpretable visual-textual correlations from them. In the following discussion, for simplicity, we deprecate the subscript i and j in V_i and Q_j to introduce the model design in terms of a single target triplet $\big(Q, V, (S, E)\big)$ where the query sentence Q and the video moment $\{\boldsymbol{x}_i^f \mid i \in [S, E] \text{ and } \boldsymbol{x}_i^f \in V\}$ are consistent in semantics.

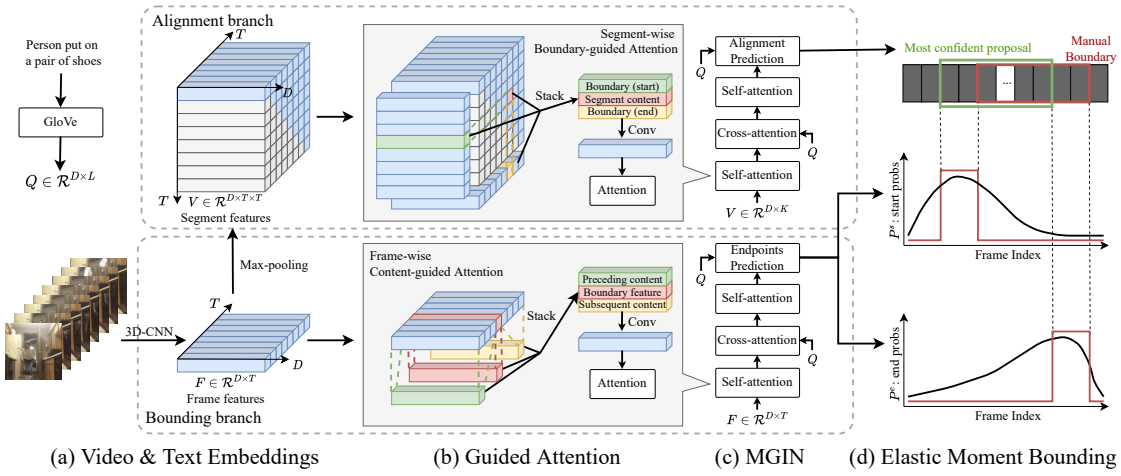


Figure 6.5: An overview of the proposed Elastic Moment Bounding method.

Approach overview. In this section, we propose an *Elastic Moment Bounding* model whose overview is depicted in Figure 6.5). The EMB model first predicts the per-frame probabilities to be the temporal endpoints of a target moment (Figure 6.5's bottom box) by a *Multi-grained Interaction Network* (MGIN) (Figure 6.5 (c)) incorporating with a *Guided Attention* mechanism (Figure 6.5 (b)). EMB then optimises the frame-wise probabilities by mining multiple candidate endpoints beyond the manual annotated ones. The candidate endpoints are discovered by an auxiliary alignment branch shown in Figure 6.5's top box. The alignment branch explores video-text content alignment in segment-level, which is less sensitive to exact endpoints annotations so more robust to endpoints uncertainty. By doing so, we construct an *elastic boundary* interpretable universally for semantically similar activities with endpoints uncertainty. In our joint content-boundary learning design, a video V will be concurrently represented in frame $\{\boldsymbol{x}_i^f\}_{i=1}^{L^f}$ and segment-level $\{\boldsymbol{x}_i^s\}_{i=1}^{L^s}$. For clarity concern, we denote video frames by V^f and video segments (proposals) by V^s in the rest of the section.

6.3.2 Temporal Endpoints Identification

Our elastic moment bounding is a generic formulation deployable in any multi-modal backbone deep networks. Here, we start with the VSLNet (Zhang et al. [217]) and reconstruct it by introducing a Guided Attention mechanism to form a multi-grained interaction network. The overall pipeline of MGIN is shown in Figure 6.5 (c) to first encode the video V^f and the sentence $Q = \{\boldsymbol{x}_i^w\}_{i=1}^{L^w}$ by attention both within (self) and across modalities, then predict the frame-wise endpoint probabilities by the joint-modal representations fused by context-query attention (Seo

et al. [159]; Xiong et al. [202]; Yu et al. [211]).

Adopting the convention (Zhang et al. [217]; Nan et al. [124]; Wang et al. [186]), video frames $V^f = \{\boldsymbol{x}_i^f\}_{i=1}^{L^f} \in \mathbb{R}^{D^v \times L^f}$ are represented by a pre-trained 3D-CNN model (Carreira and Zisserman [18]) and the query sentence by the GloVe embeddings (Pennington et al. [140]) of words $Q = \{\boldsymbol{x}_i^w\}_{i=1}^{L^w} \in \mathbb{R}^{D^t \times L^w}$. To facilitate cross-modal feature interactions, we map both the representations to have the same dimension D by two independent linear projections, *i.e.*, $V^f \leftarrow \text{FC}(V^f) \in \mathbb{R}^{D \times L^f}$ and $Q \leftarrow \text{FC}(Q) \in \mathbb{R}^{D \times L^w}$.

Vision-language attention representation. Similar as in CRM (section 6.1.2), we deploy attentive encoding (Vaswani et al. [182]) for both the visual and textual representations to explore the dependencies among elements in both. We reiterate the formulations here for clarity. In general, to encode a target sequence $X^t \in \mathbb{R}^{D \times L^t}$ of L^t elements with the help of a reference sequence $X^r \in \mathbb{R}^{D \times L^r}$ in size L^r, we first compute an attention matrix $\mathcal{A}$ indicating the pairwise target-reference correlations, then represent each target element by its correlated references:

$$\mathcal{A} = \text{FC}(X^t)^\top \text{FC}(X^r)/\sqrt{D} \in \mathbb{R}^{L^t \times L^r} \tag{6.12}$$

$$\delta(X^t, X^r) = \text{FC}(X^t + \text{FC}(X^r)\text{Softmax}(\mathcal{A})^\top) \in \mathbb{R}^{D \times L^t}. \tag{6.13}$$

An attention layer formulated in Eq. (6.13) is parameterised by three independent fully-connected layers with the same input and output channels. Our MGIN shown in Figure 6.5 (c) is constructed by both self-attention within modalities: $V^f \leftarrow \delta(V^f, V^f)$, $Q \leftarrow \delta(Q, Q)$ for context exploration and cross-attention between modalities: $V^f \leftarrow \delta(V^f, Q)$, $Q \leftarrow \delta(Q, V^f)$ to learn the semantic correlations between video frames and query words.

Guided attention. To effectively locate the temporal endpoints of activities, it is essential for the model to be aware of not only what is shown in each individual frame but also what's changing across it, *i.e.*, the differences between its preceding and subsequent content. As a simple example, the starting point of an activity 'person puts on shoes' should not be arbitrary frames involving shoes-like objects in-between the period but the one when the shoes first appear to interact with a person at the beginning of the process. Therefore, we propose a content-guided attention module (Figure 6.5 (b)'s bottom) to explicitly encode the preceding and subsequent content information

of each frame into its representation:

$$\begin{aligned} V^f_{\text{pre}} &= \{\text{MaxPool}(\{\boldsymbol{x}^f_i\}_{i=1}^{j})\}_{j=1}^{L^f} \in \mathbb{R}^{D\times L^f}, \\ V^f_{\text{sub}} &= \{\text{MaxPool}(\{\boldsymbol{x}^f_i\}_{i=j}^{L^f})\}_{j=1}^{L^f} \in \mathbb{R}^{D\times L^f}, \\ \tilde{V}^f &= \text{Conv2d}(\{V^f, V^f_{\text{pre}}, V^f_{\text{sub}}\}) \in \mathbb{R}^{D\times L^f}. \end{aligned} \tag{6.14}$$

The feature $\text{MaxPool}(\{\boldsymbol{x}^f_i\}_{i=1}^{j}) \in \mathbb{R}^D$ in Eq. 6.14 aggregate all the frames before $\boldsymbol{x}^f_j$ by max-pooling as its preceding content representation. Similarly, the subsequent content of the j-th frame is obtained by $\text{MaxPool}(\{\boldsymbol{x}^f_i\}_{i=j}^{L^f})$. Both the preceding V^f_{pre} and subsequent V^f_{sub} content features are then stacked and assembled with the frame-wise representations V^f by a 2D convolution layer. After that, the content-guided representations of video frames $\tilde{V}^f$ are used for attentive encoding (Eq. (6.13)) both within $V^f \leftarrow \delta(\tilde{V}^f, \tilde{V}^f)$ and across modalities $V^f \leftarrow \delta(\tilde{V}^f, Q)$.

Boundary prediction. Given a video $V^f \in \mathbb{R}^{D\times L^f}$ and sentence $Q \in \mathbb{R}^{D\times L^w}$ representations, we estimate the frame-wise endpoint probabilities by computing context-query attention (Seo et al. [159]; Xiong et al. [202]; Yu et al. [211]), same as the baseline (Zhang et al. [217]). It is defined as:

$$\begin{aligned} (\boldsymbol{p}^s, \boldsymbol{p}^e) &= \text{Softmax}(\text{LSTM}(\hat{V}^f \odot \boldsymbol{h})), \text{where } \boldsymbol{h} = \sigma(\text{Conv1d}(\hat{V}^f \| \boldsymbol{x}^q)) \in \mathbb{R}^{1\times L^f}, \\ \hat{V}^f &= g(V^f, Q) = \text{FC}(V^f \| X^{v2q} \| V^f \odot X^{v2q} \| V^f \odot X^{q2v}) \in \mathbb{R}^{D\times L^f}; \text{ and} \\ \mathcal{A} &= \frac{\text{FC}(V^f)^\top \text{FC}(Q)}{\sqrt{D}},\ X^{v2q} = Q\mathcal{A}^{r\top},\ X^{q2v} = V^f \mathcal{A}^r \mathcal{A}^{c\top}. \end{aligned} \tag{6.15}$$

In Eq. (6.15), we predict the frame-wise endpoint probabilities by two stacked LSTM with each followed by an independent linear layer. This is based on fusing video frames V^f with the query sentence Q by function $g(\cdot)$ then rescale the per-frame fused features $\hat{V}^f \in \mathbb{R}^{D\times L^f}$ using their estimated likelihood $\boldsymbol{h} \in \mathbb{R}^{1\times L^f}$ of being foreground (within the target moment) to suppress any distractions from redundant frames. Matrix $\mathcal{A} \in \mathbb{R}^{L^f \times L^w}$ consists of frame-to-word correlation scores; $\mathcal{A}^r$ and $\mathcal{A}^c$ are its row and column-wise softmax normalised copies. The $\boldsymbol{x}^q$ is the sentence-level representation obtained by weighted sum of words (Bahdanau et al. [5]). Notation $(\cdot\|\cdot)$ stands for concatenation (broadcast if necessary) while $\odot$ is the Hadamard Product.

6.3.3 Elastic Moment Bounding

Given the uncertainty and ambiguity in manually annotated activity temporal boundaries, it is ineffective to decide heuristically which frames and how many of them are possibly more universal and should be taken as the candidate endpoints $(\hat{\boldsymbol{S}}, \hat{\boldsymbol{E}})$ for different video activities. To address

this problem, we formulate an auxiliary alignment branch in the model to learn the video-text content mapping per each video segment. It serves as an additional self-learning "annotator" to expand the given single pair of manually annotated boundaries into candidate endpoints proposal sets tailored for individual activities.

Elastic boundary construction. As shown in Figure 6.5 (top), we first generate a 2D feature map (Zhang et al. [221]) by enumerating all pairs of frames as the start-end boundaries to represent $L^s = L^f \times L^f$ video segments $V^s = \{\boldsymbol{x}_i^s\}_{i=1}^{L^s} \in \mathbb{R}^{D\times L^s}$ as the proposals for a target moment. We flatten the 2D map here so that video segments V^s are in consistent dimensions as that of frames V^f to be encoded by MGIN. The i-th proposal with the temporal boundary of (t_i^s, t_i^e) is represented by max-pooling the frames it is composed of $\boldsymbol{x}_i^s = \text{MaxPool}(\{\boldsymbol{x}_j^f | \forall j \in [t_i^s, t_i^e]\})$. The segment-wise representations will then be fed into an independent MGIN equipped with *boundary-guided* attention modules (Figure 6.5 (b)'s top box) for visual encoding. Similar as in the *content-guided* attention for video frames, we explicitly assemble the frame-wise boundary features with the content representations of video segments to encourage boundary-sensitive content alignment:

$$\begin{aligned} V_{\text{sta}}^s = \{\boldsymbol{x}_{t_j^s}^f\}_{j=1}^{L^s} \in \mathbb{R}^{D\times L^s},\ V_{\text{end}}^s = \{\boldsymbol{x}_{t_j^e}^f\}_{j=1}^{L^s} \in \mathbb{R}^{D\times L^s}, \\ \tilde{V}^s = \text{Conv2d}(\{V^s, V_{\text{sta}}^s, V_{\text{end}}^s\}) \in \mathbb{R}^{D\times L^s}. \end{aligned} \tag{6.16}$$

The features V_{sta}^s and V_{end}^s in Eq. (6.16) are the representations of the start and end frames for each of the L^s proposals. They are stacked and assembled with the segment-wise content features V^s to derive the boundary guided segment representations $\tilde{V}^s$ by a 2D convolution layer. Such boundary-guided attention share a similar spirit with temporal pyramid pooling (Zhao et al. [227]), that is to explicitly encode the temporal structure into segment's representation so to be sensitive to its boundary. $\tilde{V}^s$ is then used for attentive encoding (Eq. (6.13)) within $V^s \leftarrow \delta(\tilde{V}^s, \tilde{V}^s)$ and across $V^s \leftarrow \delta(\tilde{V}^s, Q)$ modalities.

Given the segment-level video representations V^s, we fuse them with the sentence features by function $g(\cdot)$ defined in Eq. (6.15), then re-arrange it to be a 2D feature map and predict the per-proposal alignment scores by a 2D convolution layer with consideration of segments' overlap:

$$\boldsymbol{p}^a = \sigma(\text{Conv2d}(g(V^s, Q))) \ s.t.\ p_i^a \in (0,1)\ \forall i \in [1, L^s]. \tag{6.17}$$

The segment-wise alignment scores $\boldsymbol{p}^a$ activated by the Sigmoid function σ is then flatten and

supervised by the temporal overlaps between every proposal and the manual boundary:

$$\begin{aligned} \alpha_i &= \text{IoU}((t_i^s, t_i^e), (S, E)) \\ y_i^a &= \begin{cases} 1, & \text{if } \alpha_i \geq \tau_u \\ 0, & \text{if } \alpha_i < \tau_l \\ \alpha_i, & \text{otherwise} \end{cases} \\ \mathcal{L}_{\text{align}}\big(V^s, Q, (S, E)\big) &= \text{BCE}(\boldsymbol{y}^a, \boldsymbol{p}^a). \end{aligned} \tag{6.18}$$

The notations τ_u and τ_l are the upper and lower overlap thresholds to control the flexibility of video-text alignment, which are set to 0.7 and 0.3 respectively as in (Zhang et al. [221]). With the learned segment-wise alignment scores $\boldsymbol{p}^a$, we take the boundary (t_*^s, t_*^e) of the most confident proposal with the greatest predicted score $p_*^a \geq p_i^a \ \forall i \in [1, L^s]$ as the pseudo boundary and construct the corresponding candidate endpoint sets by:

$$\hat{\boldsymbol{S}} = [\min(t_*^s, S), \max(t_*^s, S)], \quad \hat{\boldsymbol{E}} = [\min(t_*^e, E), \max(t_*^e, E)]. \tag{6.19}$$

We customise the candidate endpoint sets for every individual activity by exploring the content alignments between video segments and query sentences, *i.e.*, elastic boundary. This is potentially more flexible and reliable than applying label smoothing globally (Wang et al. [186]; Xiao et al. [199]) without considering video context and language semantics.

Reliability *vs*. flexibility. Introducing too many candidate endpoints that are semantically irrelevant to the query sentences is prone to distracting the model from learning effective visual-textual correlations, especially at the early stage of training a randomly initialised model which is likely to yield inaccurate pseudo boundaries (t_*^s, t_*^e). Therefore, we balance the reliability and flexibility of our elastic boundary by a controllable threshold τ:

$$t_*^s, t_*^e = \arg\max_{t_i^s, t_i^e} \boldsymbol{p}^a \quad s.t. \quad \alpha_i \geq \tau. \tag{6.20}$$

The α_i in Eq. (6.20) implies the overlap between the i-th proposal and the manual boundary, whilst the threshold τ serving as a controllable trade-off between flexibility and reliability so that only the sufficiently overlapped proposals will be selected for constructing the elastic boundary in Eq. (6.19).

Learning from elastic boundary. With the elastic boundary $(\hat{\boldsymbol{S}}, \hat{\boldsymbol{E}})$, instead of optimising model's predictions (Eq. (6.15)) to be exactly the same as the rigid manual endpoint label (S, E),

we construct the frame-wise supervisions to maximise the sum of candidate's probabilities:

$$\mathcal{L}_{\text{bound}}\big(V^f, Q, (S,E)\big) = -\log(\sum_{i\in\hat{\boldsymbol{S}}} p_i^s) - \log(\sum_{i\in\hat{\boldsymbol{E}}} p_i^e). \tag{6.21}$$

We optimise the sum rather than product of probabilities to allow the model to select confident endpoints from elastic boundary instead of encouraging uniform predictions for all the candidate frames. Comparing with the commonly adopted frame-wise supervision which trains $\boldsymbol{p}^s$ and $\boldsymbol{p}^e$ to be one-hot (Zhang et al. [217]; Nan et al. [124]), we provide in Eq. (6.21) a more flexible boundary to the target moments so that the model can learn in a data-driven manner to select the endpoints beyond the manual boundary and ignore the unconcerned actions involved.

6.3.4 Model Training and Inference

Inference. We consider two schemes when predicting the boundary of video activity: **(a)** following the standard protocol of the task (Gao et al. [50]; Krishna et al. [91]), we predict a determined (DET) boundary enclosed by a single start and end frames according to the outputs of bounding branch in a maximum likelihood manner:

$$\tilde{S} = \arg\max_i \boldsymbol{p}^s, \quad \tilde{E} = \arg\max_i \boldsymbol{p}^e, \tag{6.22}$$

where $\tilde{S}$ and $\tilde{E}$ are the predicted start and end frame indices of a video that are corresponding to a given query. **(b)** Besides, considering the uncertain nature of temporal boundary, it is more intuitive to estimate the endpoints of video activity by temporal spans rather than specific frames. Our model is able to predict also an elastic (ELA) boundary in a similar way as in training

$$\tilde{\boldsymbol{S}} = [\min(t_*^s, \tilde{S}), \max(t_*^s, \tilde{S})], \quad \tilde{\boldsymbol{E}} = [\min(t_*^e, \tilde{E}), \max(t_*^e, \tilde{E})]. \tag{6.23}$$

In Eq. (6.23), we denote $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{E}}$ in bold to indicate a set of candidate endpoints, and differentiate them from the determined boundary in Eq. (6.22). The (t_*^s, t_*^e) is the boundary of the most confident proposals selected from the alignment branch without constraint on their overlaps to the manual boundary. *i.e.*, $\tau = 0$ in Eq. (6.20).

Training. The bounding and alignment branches in the EMB model are simultaneously trained in the conventional batch-wise manner. A mini-batch is composed of n_{bs} pairs of untrimmed video and natural language sentence, each with a manual boundary (S,E) indicating the video's interval corresponding to the query, *i.e.*, the target activity moment. The overall loss function for a video-query-boundary triplet is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{bound}} + \lambda_2 \mathcal{L}_{\text{align}}, \tag{6.24}$$

where λ_1 and λ_2 are designed to balance the two branches, both of which are empirically set to 1 as they play equally important roles in training while neither of them numerically dominates the loss by sharing the same cross-entropy based formulation. The EMB model is optimised end-to-end by minimising $\mathcal{L}$ using stochastic gradient descent. Its overall training process is summarised in Algorithm. 6.

Algorithm 6: EMB for video grounding with uncertain labels.

Input: An untrimmed video V^f, a query sentence Q, a temporal boundary (S, E).

Output: An updated video activity localisation model.

`// Bounding branch`

Encode frames by content-guided attention via Eq. (6.13)(6.14);

Fuse frames with query and predict per-frame endpoint probabilities via Eq. (6.15);

`// Alignment branch`

Construct 2D feature map of proposals V^s;

Encode proposals by boundary-guided attention via Eq. (6.13)(6.16);

Fuse proposals with query and predict proposal-query alignment scores via Eq. (6.17);

`// Supervisions to both the branches`

Construct the elastic boundary via Eq. (6.19)(6.20);

Optimise model weights by minimising $\mathcal{L}$ in Eq. (6.24);

6.4 Experiments on Knowledge Propagation from Uncertain Labels

6.4.1 Datasets and Metrics

Datasets. Beyond the two benchmark datasets adopted in Section 6.2, we further evaluated our EMB model on the TACoS dataset to facilitate extensive comparisons with contemporary methods trained by manual labels. To be concrete, our experiments were carried out on three benchmark datasets, including **(1)** TACoS (Regneri et al. [148]) adapted from the MPII Cooking Composite Activities dataset (Rohrbach et al. [151]), **(2)** ActivityNet-Captions (Krishna et al. [91]; Heilbron et al. [70]) and **(3)** Charades-STA (Gao et al. [50]; Sigurdsson et al. [162]). We summarise their different data characteristics in Table 6.4. Among the three datasets, the raw videos in TACoS have the longest durations (287.14s) while that of its MoIs are shortest in contrast (5.45s), which means that the video activities are temporally covering less than 2% of

the complete videos on average. Therefore, the videos in TACoS contain a lot of redundancy in terms of every MoIs. On the other hand, the ActivityNet is very different from TACoS whose activities temporally cover much larger proportions of the videos ($\sim 30\%$) than the other two.

Dataset	#Video	#Train	#Val	#Test	L^v	L^m	L^w
TACoS	127	10,146	4,589	4,083	287.14s	5.45s	10.1
ActivityNet	19,290	37,421	17,031	17,505	117.61s	36.18s	14.8
Charades	6,672	12,408	-	3,720	30.59	8.22	7.2

Table 6.4: Statistics of video activity localisation datasets. L^v and L^m are the average lengths of untrimmed videos and MoIs respectively, whilst L^w is the average number of words in query sentences. The higher ratio between L^v and L^m with shorter L^w dictate harder localisation tasks.

Evaluation metrics. Similar as in Section 6.2, we measure the quality of video activity localisation results by their average recall rate at different temporal IoU thresholds (IoU@m). The predicted boundary $(\tilde{S},\tilde{E})$ of a MoI obtained by Eq. (6.22) is considered correct if its IoU with the manual temporal label (S,E) is greater than the thresholds m which are predefined as $m = \{0.3, 0.5, 0.7\}$. Besides, we also reported the Mean Intersection over Union (mIoU) of all predictions with their corresponding ground-truth temporal labels to show the average overlaps between the predicted and manual boundaries. Specially, for our elastic boundary, we enumerate all the start-end pairs from $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{E}}$ (Eq. (6.23)), respectively. If a manual boundary's overlap to any of the combinations is greater than the IoU threshold, we consider it is correctly predicted.

6.4.2 Implementation Details

For fair comparisons, we adopted the video features provided by our baseline model (Zhang et al. [217]) and the 300-D GloVe (Pennington et al. [140]) embeddings to encode the input video frames and query sentences respectively, both remaining fixed during training. We downsampled videos to have 128 frames at most by max-pooling and zero-padded the shorter ones. The output dimension of all the hidden layers are set to 128 as in (Zhang et al. [217]) and the multi-head variant (Vaswani et al. [182]) of the attention layer in Eq. (6.12) was used with 8 heads followed by layer normalisation and random dropout at 0.2. The EMB model was trained for 100 epochs with a batch size of 16. It was optimised by an Adam optimiser using a linearly decaying learning rate of 0.0005 and gradient clipping of 1.0. In the alignment branch, we downsampled the videos to have 16 clips by the max-pooling of every 8 continuous frames for constructing

the 2D feature maps of video segments to avoid over-dense proposals. For training the bounding branch, we construct the elastic boundary subject to an evolving threshold τ in Eq. (6.20), which progressively decreased from 1 to 0.5 in the course of training.

6.4.3 Comparisons with the State-of-the-Art

We evaluate the effectiveness of the proposed EMB model by comparing it with 13 state-of-the-art methods, including VSLNet (Zhang et al. [217]), IVG (Nan et al. [124]), 2D-TAN (Zhang et al. [221]), LGI (Mun et al. [123]), DPIN (Wang et al. [185]), DRN (Zeng et al. [214]), SCDM (Yuan et al. [212]), BPNet (Xiao et al. [199]), CPNet (Li et al. [105]), CPN (Zhao et al. [228]), DeNet (Zhou et al. [229]), CBLN (Liu et al. [108]), and SMIN (Wang et al. [186]) Table 6.5 shows clear performance advantages of the elastic boundary predicted by our EMB model over the contemporary competitors. When constructing the elastic boundaries in inference, over 80% of the predictions pairs yielded by the alignment and bounding branches are consistent with each other (IoU > 0.5). Therefore, the performance improvements we obtained is not due to over-dense sampling of the potential boundaries. However, to learn with elastic boundary, it takes the EMB model 4.3 minutes to train for one epoch on ActivityNet-Captions *vs*. 1.5 minutes for VSLNet (Zhang et al. [217]). The testing time also increases from 20.6 to 50.6 seconds for the construction of elastic boundary with the auxiliary alignment branch. Besides, as for the determined boundary yielded by our models, EMB (DET) outperforms the state-of-the-art methods on TACoS against all the performance metrics while remaining its competitiveness on the other two datasets. It is important to note that among the three datasets, TACoS poses the hardest test with the longest average untrimmed videos and the shortest activity moments of interest (see Table 6.4). That is, TACoS exhibits more realistic scenarios for activity localisation test. In this context, EMB shows its advantage over other models most clearly when the untrimmed videos are longer whilst the video MoIs are sparse and far between. Furthermore, the determined boundary yielded by EMB also outperforms the baseline VSLNet (Zhang et al. [217]) by significant margins on all tests. The more recent IVG (Nan et al. [124]) was also based on VSLNet, sharing the same baseline as EMB. The notable performance advantages of EMB over both VSLNet and IVG demonstrate its non-trivial improvements.

Method	TACoS				Charades-STA				ActivityNet-Captions			
	mIoU	IoU@m			mIoU	IoU@m			mIoU	IoU@m		
		0.3	0.5	0.7		0.3	0.5	0.7		0.3	0.5	0.7
VSLNet	24.11	39.61	24.27	20.03	45.15	64.30	47.41	30.19	43.19	63.13	43.22	26.16
IVG	28.26	38.84	29.07	19.05	48.02	67.63	50.24	32.88	44.21	63.22	43.83	27.10
2D-TAN	-	37.29	25.32	-	-	-	39.81	23.25	-	59.45	44.51	26.54
LGI	-	-	-	-	51.38	72.96	59.46	35.48	41.13	58.52	41.51	23.07
DPIN	-	46.74	32.92	-	-	-	47.98	26.96	-	62.40	47.27	28.31
DRN	-	-	23.17	-	-	-	53.09	31.75	-	-	45.45	24.36
SCDM	-	26.11	21.17	-	-	-	54.44	33.43	-	54.80	36.75	19.86
BPNet	19.53	25.93	20.96	14.08	46.34	65.48	50.75	31.64	42.11	58.98	32.07	24.69
CPNet	28.69	42.61	28.29	-	52.00	-	60.27	38.74	40.65	-	40.56	21.63
CPN	34.63	48.29	36.58	21.25	51.85	72.94	56.70	36.62	45.70	62.81	45.10	28.10
DeNet	-	-	-	-	-	-	59.75	38.52	-	61.93	43.79	-
CBLN	-	38.98	27.65	-	-	-	61.13	38.22	-	66.34	48.12	27.60
SMIN	-	48.01	35.24	-	-	-	64.06	40.75	-	-	48.46	30.34
VSLNet†	28.15	39.07	27.59	16.65	47.33	67.26	50.46	31.53	42.26	57.75	41.10	25.58
EMB (DET)	36.66	51.74	38.74	24.47	52.47	72.10	58.52	39.65	45.63	63.72	46.72	27.62
EMB (ELA)	**49.08**	**64.61**	**53.29**	**38.42**	**62.77**	**80.22**	**69.87**	**52.98**	**55.41**	**73.30**	**57.21**	**39.07**

Table 6.5: Performance comparisons of EMB to the state-of-the-art models on three video activity localisation benchmark datasets. The 1st/2nd best results are highlighted in **red**/blue, respectively. The 'DET' modifier of EMB stands for the determined boundary predicted in Eq. (6.22) while 'ELA' is the elastic boundary (Eq. (6.23)). The symbol † denotes the reproduced results of our baseline model under the strictly identical setups using the code from authors.

6.4.4 Component Analysis and Ablation Study

For in-depth understandings of the proposed EMB model, we conducted comprehensive ablation studies by comparing our determined predictions with the ones yielded by our baseline (Zhang et al. [217]) to investigate the effectiveness of model designs.

Components analysis. We investigated the individual contributions of different components in our EMB model to its improvements over the baseline model (Zhang et al. [217]). As shown in Figure 6.6, both our elastic boundary learning objective (Eq. (6.21)) and the MGIN brought clear benefits to the baseline. Such results demonstrate the effectiveness to learn the temporal endpoints of video activities with higher flexibility so to tolerant the uncertainty of manual labels. Besides, they also imply the superiority of our visual encoders which conduct both within and cross-modal attention learning and complement the boundary and content information of video

segments mutually.

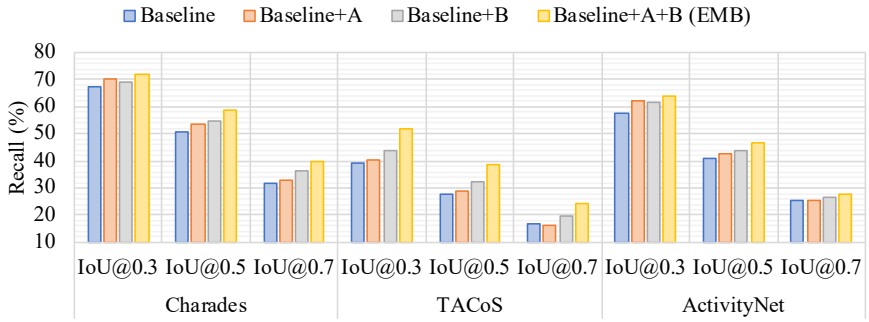


Figure 6.6: Effects of the proposed components in EMB. The elastic moment bounding formulation is denoted as component "A" while the multi-grained interaction network is "B".

Candidate endpoints mining. We evaluated the advantages of mining candidate endpoints adaptively over several heuristic strategies: (1) boundary extension (Zhang et al. [217]), (2) smoothing by a gaussian kernel (Wang et al. [186]; Xiao et al. [199]) and (3) fitting single-frame manual endpoints (baseline). As shown in Figure 6.7, simply improving the boundary's flexibility without considering their reliability ("Extend") tends to degrade the model's performances on both datasets. Boundary smoothing by a gaussian kernel ("Kernel") is sometimes beneficial but less stable than our adaptive designs. This is because their candidates were determined according to only the duration of MoIs without considering the video context and query's unambiguity.

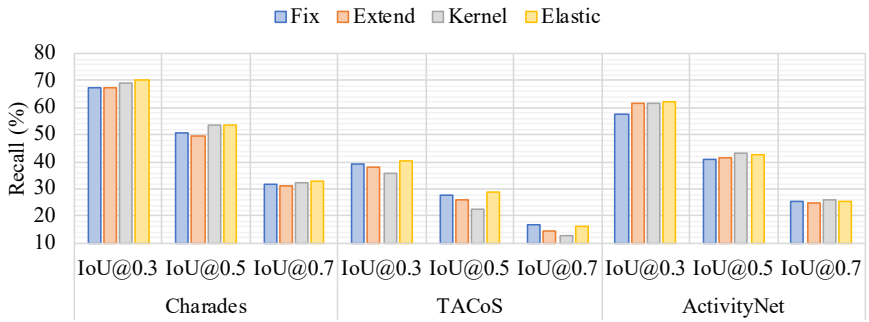


Figure 6.7: Effects of candidate endpoints mining strategies used in training. "Fix": single-frame manual boundaries. Multiple candidate endpoints are generated by extension ("Extend"), a gaussian kernel ("Kernel"), or our elastic bounding.

Evolving threshold. We studied the effects of threshold's evolving schemes to our elastic boundary constructions (Eq. (6.20)). Figure 6.8 shows the curves of schemes and their corresponding performances on Charades-STA. The model trained with a constant threshold yielded the worst results in most cases, which indicates that an evolving threshold is beneficial as a trade-off between reliability and flexibility at different training stages. Besides, the model learned

with the "Sigmoid" scheme is superior to others in all cases. By comparing the curves in Figure 6.8, it's clear that the 'Sigmoid' scheme maintains a persistently high threshold at the early training stages to avoid introducing distractions to boundary identification when the alignment branch is under-trained, then drops rapidly to involve more diverse candidate endpoints when the alignment branch is reliable.

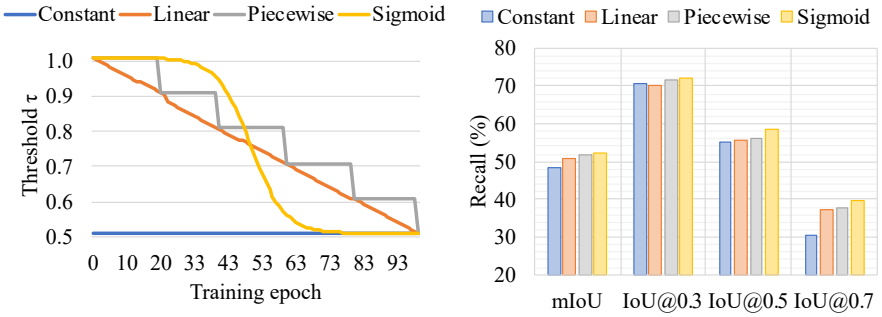


Figure 6.8: Effect of constructing elastic boundary subject to an evolving threshold on Charades.

Guided attention. We validated the effectiveness of our guided attention mechanism by replacing it in our MGIN encoder by the conventional attention modules proposed in (Vaswani et al. [182]). From the comparison results shown in Figure 6.9, the models trained with guided attention outperformed their counterparts which learned the video representations without interacting information in multiple granularities. Such results imply the complement of segment's content and boundary information, which encourages the video feature representations to be sensitive to redundancy and activity transitions.

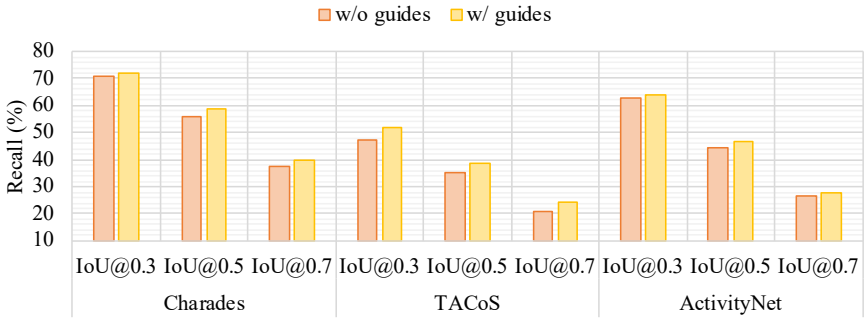


Figure 6.9: Effect of guided attention mechanism by comparing with the conventional attention modules (Vaswani et al. [182]).

Qualitative case study. For qualitative study, we provide several video examples in Figure 6.10 which are showing video activities corresponding to semantically similar sentence descriptions.

However, their manual boundary are inconsistent, demonstrating the uncertainty in temporal boundary. Specifically, the manual boundary for Q_1 starts from grabbing the food right before putting it into the mouth while Q_2 skipping the action of "grabbing" and start when the person takes a bite. The Q_3 involves even more redundancy which covers the actions of taking a plate from a desk and blending foods by a folk, before the person really deliver the food to his mouth. In contrast, the predictions made by our model are more consistent on interpreting the action of "eat" in different videos, *i.e.*, always starts from delivering food to the mouth. This is because, by reliably exploring multiple candidate endpoints for the target moments, we train the model to locate video activities in higher flexibility so that the model can learn from the boundary that are universally interpretable instead of fitting the pre-fixed manual endpoints which will lead to visual-textual mis-correlations.

Figure 6.10: Cases of video activities on ActivityNet-Captions with similar semantics but inconsistent manual boundaries. The manual and predicted boundary are shown in red and green, respectively.

6.5 Summary

This chapter explores knowledge propagation from incomplete labels to learn discriminative representations for video segments, so as to accurately locate moments-of-interest in untrimmed videos. Specifically, we focus on two common challenges when the incomplete labels are given in *collective* or *unreliable*. In the case of learning from collective labels, we presented a novel *Cross-sentence Relations Mining* method for learning video activity localisation in the absence of per-sentence temporal annotation, *i.e.* the descriptions of activity are associated with videos rather than segments. CRM explores cross-sentence relations within each paragraph descrip-

tion of a long video to optimise video moment proposal selections in training so to improve per-sentence localisation in testing. CRM minimises mismatching individual sentences to video moment proposals during training by constraining their selections according to the temporal ordering and pairwise sentencing as expanded queries in the context of a paragraph description of video. This improves notably CRM's capacity to locate more accurately video activities against more complex language descriptions. For learning from unreliable labels, we introduced a new *Elastic Moment Bounding* approach to learn a more robust model by constructing elastic boundary tailored to learning more flexibly the endpoints of every target moment. The EMB is enabled and encouraged to select the optimal boundary for each video moments by exploring semantically similar activities, rather than being enforced to predict the rigid manual labels. With such an elastic bounding design, EMB learns a more accurate and universal visual-textual correlation generalisable to activity moment localisation in more naturally prolonged unseen videos where activity of interests are fractionally small and harder to detect. Comprehensive experiments on three activity localisation benchmark datasets demonstrate the competitiveness and unique advantages of both the CRM and EMB models over the state-of-the-art models. Extensive ablation studies further provided in-depth analysis and understandings of the individual components formulated in them.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis presented a comprehensive study on unsupervised deep learning of visual representations. Particularly, the conventional 'pure' unsupervised learning without any manual labels was expanded to a more generalised setup to discuss deep visual feature learning when human annotations are *insufficient and/or inadequate* for constructing reliable supervisions to underlie the mapping functions from data to the desired predictions. In this context, four types of generalised unsupervised learning problems that widely exist in various computer vision applications were investigated, including knowledge aggregation from local data structure, knowledge discovering from global data structure, knowledge transferring from relevant labels, and knowledge propagation from incomplete labels. Whilst remarkable progress have been made by recent works in the literatures, there are still some remaining challenges to be resolved. To this end, novel methods were proposed in this thesis to address the limitations of existing solutions. Specifically,

(1) In Chapter 3, a novel progressive neighbourhood mining idea is introduced to learn discriminative representations of *unlabelled* imagery data by exploring inter-sample adjacent relationships in local neighbourhoods with a divide-and-conquer principle, which combines the merits of cluster analysis and sample specificity learning while mitigating their disadvantages in being misled by inaccurate pseudo cluster labels or insensitive to within-class visual variations. As an instantiation, an Anchor Neighbourhood Discovery method is proposed to progressively learn with confidently positive sample pairs in a curriculum

learning scheme to improve model's robustness to unreliable positive signals which are inconsistent with the underlying class memberships. Furthermore, to better balance the size of local neighbourhoods and their purity (*i.e.*, true positive rate), a novel Progressive Affinity Diffusion model is introduced to broadcast the potential positive inter-sample relationships across adjacent neighbourhoods with a cyclic constraint. By expanding local neighbourhoods while maintaining their consistency with underlying class memberships, PAD is sensitive to complexer intra-class visual variations and higher inter-class similarity, hence, more class discriminative than training with neighbourhoods in restricted size.

(2) In Chapter 4, class discriminative visual representations are learned from the global cluster structure *without any human labels* by two new deep clustering methods PartItion Confidence mAximisation and Semantic Contrastive Learning. Specifically, the PICA model is formulated to derive the decision boundary of the underlying semantic classes by maximising the partition confidence which is beneficial to improve the semantic plausibility of the resulted clusters. On the other hand, SCL is a novel variant of instance contrastive learning to encourage the sample specificity learning to be sensitive to underlying class memberships by imposing explicitly distance-based cluster structures on unlabelled training data to enable cross-cluster instance discrimination. Inheriting the class-insensitive formulation of sample specificity learning, the SCL model is robust to the errors propagated from inaccurate cluster assignments to visual feature learning.

(3) In Chapter 5, a new self-SUPervised REMEdy approach is proposed to facilitate effective knowledge transfer from a *relevant labelled* source domain to benefit clustering of images on an unlabelled target domain. Considering that the relationships between different data domains are usually complex and unpredictable in practice, the SUPREME model makes no assumption on seen target data as in FSL or known target classes in relation to the source classes as in ZSL and domain adaptation. In this case, SUPREME complements the insufficient transferred supervisions constructed by non-transferable knowledge acquired in source domains by self-supervisions intrinsically available in target domains and adaptively balances the two types of supervision on each independent instances to ensure learning discriminative visual representations from the hard samples where the transferred supervisions are ambiguous.

(4) In Chapter 6, a novel Cross-sentence Relations Mining model for video activity localisation by natural language is developed to acquire accurate understandings of video-text alignment from human annotations on associating descriptions with videos rather than video moments, *i.e*. incomplete labels given *in collective*. The CRM model exploits the cross-sentence relationships in a video's paragraph description to explicitly constrain cross-moment relationships, so as to alleviate the ambiguity of each individual sentence. Despite such constraints are less complete than the supervisions constructed by per-sentence fine-grained temporal annotations, they are more practical to be collected for model training. Moreover, a new Elastic Moment Bounding method is proposed in Chapter 6 for video activity localisation with inherently *uncertain* temporal labels which will lead to ambiguous supervision signals when being considered as oracles, and result in biased visual-textual correlation. In EMB, each video moments is modelled adaptively by an elastic boundary with a set of candidate endpoints. This enables the model to learn to select optimal endpoints from consistent visual-textual correlations among semantically similar activities and obtains universal knowledge about video-text alignments.

7.2 Future Work

Beyond the methods developed in this thesis, there are still some remaining limitations and challenges in the discussed problems that are worth to be further investigated in future works. These are summarised as follows:

(1) **Aggregating knowledge from local data structure in representation learning**: The AND and PAD models proposed in this thesis yielded promising performances on learning discriminative visual features without any human annotations, whilst the benchmark datasets widely adopted in the field are mostly well-curated to be object-centric (Locatello et al. [111]; Russakovsky et al. [153]), *i.e*., the objects of interest occupy the main and most salient regions. However, large-scale natural images collected in the wild are usually with complex background and in arbitrary structures. Therefore, effective attention learning without constraints from manual labels is worthy of further study, especially with the help of the rising transformer-like models (Vaswani et al. [182]; Devlin et al. [38]) which have been shown remarkable on learning both sequential and spatial attention.

(2) **Discovering knowledge from global data structure in deep clustering**: Whilst the two

proposed PICA and SCL approaches are able to derive discriminative visual features and semantically plausible separations for unlabelled image data, they hold two stringent assumptions which are common in the literature of deep clustering but not always valid in practice. First, the number of clusters are trivially assumed known. It is challenging to accurately estimate how many classes are involved in a collection of unlabelled images without labelling and knowing their semantics. Although this is a long-standing and critical problems in clustering (Wang et al. [187]; Milligan and Cooper [120]; Maulik and Bandyopadhyay [117]), it will be interesting and beneficial to further aggregate such an estimation task into the end-to-end learning of clusters and visual features to approach to their global optimum. Another assumption is on the balanced class distributions. Such an assumption is usually made to avoid trivial clustering solutions when a majority of samples are assigned into minor clusters (Xie et al. [200]; Yang et al. [206]). However, it is well-known that natural data usually follows a long-tailed distribution (Dong et al. [42]; Japkowicz and Stephen [83]; Weiss [191]), hence, how to deal with class imbalance in deep clustering remains an open question.

(3) Transferring knowledge from relevant labels in transfer clustering: The SUPREME model presented in this thesis is formulated with fewer assumptions on the unlabelled target data and their underlying semantic classes, leading to a more scalable model which is applicable in arbitrary transfer learning scenarios. However, it is interesting to conduct further study on how can such a generic model benefits more transfer learning tasks beyond direct deployment when additional information is available, *e.g*. fine-tuning or by more advanced designs to make use of extra expert knowledge. For example, when the target classes are known in relation to source classes in zero-shot learning, it is possible to obtain the initial cluster assignments according to such knowledge without relying on independent clustering algorithms like k-means.

(4) Propagating knowledge from incomplete labels in video activity localisation: Although the presented CRM and EMB models are designed to learn universal visual-textual correlations from collective or uncertain human annotations, the obtained video-text knowledge is likely biased to the training data given its restricted scale. Considering that an untrimmed and unstructured video is equivalent to a sequence of hundreds or thousands of images, collecting and curating large-scale video data is fundamentally a more challenging task

than labelling individual images in similar scale, let alone making annotations. Given such biased knowledge, one drawback of existing localisation models is their poor scalability to locate video activity in videos filmed in novel scenarios and described by unknown vocabulary (Bao and Mu [6]). Along with the rapid developments of the image-text foundation models (Radford et al. [144]; Qi et al. [141]) which acquires comprehensive visual-textual knowledge from massive amount of pairwise image-text data, one potential solution will be bridging videos and text through images to facilitate more generic localisation models which is applicable to unseen scenes for unknown activities.

Bibliography

[1] R. Achanta and S. Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4660, 2017.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38 (7):1425–1438, 2015.

[3] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017.

[4] Y. M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[6] P. Bao and Y. Mu. Learning sample importance for cross-scenario video temporal grounding. *arXiv preprint arXiv:2201.02848*, 2022.

[7] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

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

[9] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Pro-*

ceedings of ICML workshop on unsupervised and transfer learning, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.

[10] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 153–160, 2007.

[11] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the International Conference on machine learning (ICML)*, pages 41–48, 2009.

[12] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013.

[13] P. Bojanowski and A. Joulin. Unsupervised learning by predicting noise. In *Proceedings of the International Conference on machine learning (ICML)*, pages 1–10, 2017.

[14] D. Cai, X. He, X. Wang, H. Bao, and J. Han. Locality preserving nonnegative matrix factorization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[15] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18, 2018.

[16] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020.

[17] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[18] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[19] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[20] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5879–5887, 2017.

[21] J. Chang, Y. Guo, L. Wang, G. Meng, S. Xiang, and C. Pan. Deep discriminative clustering analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–11, 2019.

[22] X. Chang, Y. Yang, T. Xiang, and T. M. Hospedales. Disjoint label space transfer learning with common factorised space. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[23] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68. Springer, 2016.

[24] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[25] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua. Temporally grounding natural sentence in video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1015. URL `https://www.aclweb.org/anthology/D18-1015`.

[26] S. Chen and Y.-G. Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–618. Springer, 2020.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[28] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

[29] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[30] Y. Chen, M. Chen, R. Wu, J. Zhu, Z. Zhu, Q. Gu, and H. Robotics. Refinement of boundary regression using uncertainty in temporal action localization. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[31] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[32] A. Coates and A. Y. Ng. Selecting receptive fields in deep networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 2528–2536, 2011.

[33] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 215–223, 2011.

[34] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[35] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[36] P. Dayan, M. Sahani, and G. Deback. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, pages 857–859, 1999.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[39] K. Do, T. Tran, and S. Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9928–9938, 2021.

[40] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by

context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[41] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[42] Q. Dong, S. Gong, and X. Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(6):1367–1381, 2018.

[43] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang. Weakly supervised dense event captioning in videos. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 3059–3069, 2018.

[44] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[45] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(11): 2765–2781, 2013.

[46] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[47] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018.

[48] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(2):303–316, 2013.

[49] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on machine learning (ICML)*, 2015.

[50] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017.

[51] M. Gao, L. S. Davis, R. Socher, and C. Xiong. Wslln: Weakly supervised natural language localization networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[52] R. Ge, J. Gao, K. Chen, and R. Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253. IEEE, 2019.

[53] Z. Ghahramani. Unsupervised learning. In *Advanced lectures on machine learning*, pages 72–112. Springer, 2004.

[54] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N19-1198`.

[55] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–16, 2018.

[56] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov. Neighbourhood components analysis. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 513–520, 2005.

[57] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[58] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[59] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[60] K. C. Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition (PR)*, 10(2):105–112, 1978.

[61] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[62] X. Guo, L. Gao, X. Liu, and J. Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1753–1759, 2017.

[63] P. Haeusser, J. Plapp, V. Golkov, E. Aljalbout, and D. Cremers. Associative deep clustering: Training a classification network with no labels. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, pages 18–32. Springer, 2018.

[64] K. Han, A. Vedaldi, and A. Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8401–8409, 2019.

[65] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[66] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[67] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[68] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[69] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[70] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

[71] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018.

[72] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[73] G. Hinton and T. J. Sejnowski. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[74] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[75] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[76] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[77] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[78] R. V. Hogg, J. McKean, and A. T. Craig. *Introduction to mathematical statistics*. Pearson Education, 2005.

[79] Y.-C. Hsu, Z. Lv, and Z. Kira. Learning to cluster in order to transfer across domains and tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–10, 2017.

[80] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira. Multi-class classification without multi-class labels. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–10, 2019.

[81] G. Huang, H. Larochelle, and S. Lacoste-Julien. Centroid networks for few-shot clustering and unsupervised few-shot classification. *arXiv preprint arXiv:1902.08605*, 2019.

[82] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on machine learning (ICML)*, pages 448–456. PMLR, 2015.

[83] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[84] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid. Deep subspace clustering networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 24–33, 2017.

[85] X. Ji, J. F. Henriques, and A. Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9865–9874, 2019.

[86] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1943–1950. IEEE, 2010.

[87] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[88] J. D. Keeler, D. E. Rumelhart, and W. K. Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563, 1991.

[89] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[90] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[91] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[94] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[95] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4281–4289, 2018.

[96] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 577–593, 2016.

[97] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *Stanford University: CS 231N Lecture notes*, pages 1–6, 2015. Accessed: 2022-6-24.

[98] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

[99] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[100] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the International Conference on machine learning (ICML)*, pages 609–616, 2009.

[101] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10657–10665, 2019.

[102] S. Lee, D. Kim, N. Kim, and S.-G. Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 91–100, 2019.

[103] J. Lei, L. Yu, T. L. Berg, and M. Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[104] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2019.

[105] K. Li, D. Guo, and M. Wang. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 1902–1910, 2021.

[106] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[107] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11539–11546, 2020.

[108] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng, W. Wei, Z. Xu, and Y. Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11235–11244, 2021.

[109] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

[110] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[111] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-centric learning with slot attention. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 33:11525–11538, 2020.

[112] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[113] D. G. Lowe et al. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 99, pages 1150–1157, 1999.

[114] J. M. Lucas and M. S. Saccucci. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12, 1990.

[115] M. Ma, S. Yoon, J. Kim, Y. Lee, S. Kang, and C. D. Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 156–171. Springer, 2020.

[116] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.

[117] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(12):1650–1654, 2002.

[118] E. R. Mayu Otani, Yuta Nakahima and J. Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[119] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 464–471. IEEE, 2000.

[120] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[121] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11592–11601, 2019.

[122] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the International Conference on machine learning (ICML)*, pages 10–18. PMLR, 2013.

[123] J. Mun, M. Cho, and B. Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10810–10819, 2020.

[124] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775, 2021.

[125] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, number 2, page 5, 2011.

[126] A. Ng. Sparse autoencoder, 2011. 2022-6-24.

[127] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 14, 2001.

[128] F. Nielsen. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer, 2016.

[129] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[130] C. Niu, J. Zhang, G. Wang, and J. Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[131] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–84, 2016.

[132] M. Noroozi, H. Pirsiavash, and P. Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2017.

[133] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):971–987, 2002.

[134] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2009.

[135] N. Patricia and B. Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1442–1449, 2014.

[136] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2751–2758. IEEE, 2012.

[137] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi. Deep subspace clustering with sparsity prior. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1925–1931, 2016.

[138] X. Peng, J. Feng, J. Lu, W.-Y. Yau, and Z. Yi. Cascade subspace clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[139] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019.

[140] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[141] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[142] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3603–3612, 2019.

[143] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[144] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021.

[145] V. V. Raghavan and C. Yu. A comparison of the stability characteristics of some graph theoretic clustering methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (4):393–402, 1981.

[146] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 876–889. Springer, 2012.

[147] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[148] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[149] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[150] D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[151] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 144–157. Springer, 2012.

[152] L. Rokach and O. Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[153] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–15. Springer, 2012.

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

[155] S. Russell and P. Norvig. Artificial intelligence: A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, 90:33–48, 2002.

[156] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, 2018.

[157] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 41–48, 2004.

[158] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2017.

[159] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[160] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[161] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8):888–905, 2000.

[162] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526. Springer, 2016.

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

[164] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song. Transductive unbiased embedding for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1024–1033, 2018.

[165] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Springer US, 1993. ISBN 978-1-4899-3216-7. doi: 10.1007/978-1-4899-3216-7_1. URL http://dx.doi.org/10.1007/978-1-4899-3216-7_1.

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

[167] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[168] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018.

[169] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[170] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[171] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[172] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[173] R. Tan, H. Xu, K. Saenko, and B. A. Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2083–2092, 2020.

[174] Y. Tang, R. Salakhutdinov, and G. Hinton. Robust boltzmann machines for recognition and denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2264–2271, 2012.

[175] Y. Tao, K. Takagi, and K. Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[176] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1 (2):146–160, 1972.

[177] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[178] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[179] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017.

[180] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[181] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–285. Springer, 2020.

[182] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[183] V. K. Verma and P. Rai. A simple exponential family framework for zero-shot learning. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 792–808. Springer, 2017.

[184] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research (JMLR)*, 11(Dec):3371–3408, 2010.

[185] H. Wang, Z.-J. Zha, X. Chen, Z. Xiong, and J. Luo. Dual path interaction network for video moment localization. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4116–4124, 2020.

[186] H. Wang, Z.-J. Zha, L. Li, D. Liu, and J. Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7026–7035, 2021.

[187] L. Wang, C. Leckie, K. Ramamohanarao, and J. Bezdek. Automatically determining the number of clusters in unlabeled data sets. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 21(3):335–350, 2009.

[188] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.

[189] X. Wang, K. He, and A. Gupta. Transitive invariance for selfsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[190] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12836–12845, 2020.

[191] G. M. Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004.

[192] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[193] J. Wu, K. Long, F. Wang, C. Qian, C. Li, Z. Lin, and H. Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–12, 2019.

[194] J. Wu, G. Li, X. Han, and L. Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1283–1291, 2020.

[195] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[196] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(9):2251–2265, 2018.

[197] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5542–5551, 2018.

[198] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10275–10284, 2019.

[199] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, and J. Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 2986–2994, 2021.

[200] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the International Conference on machine learning (ICML)*, pages 478–487, 2016.

[201] T.-T. Xie, C. Tzelepis, and I. Patras. Boundary uncertainty in a single-stage temporal action localization network. *arXiv preprint arXiv:2008.11170*, 2020.

[202] C. Xiong, V. Zhong, and R. Socher. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*, 2016.

[203] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 1537–1544, 2005.

[204] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[205] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[206] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the International Conference on machine learning (ICML)*, pages 3861–3870. JMLR. org, 2017.

[207] J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5147–5156, 2016.

[208] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha. Learning classifier synthesis for generalized few-shot learning. *arXiv preprint arXiv:1906.02944*, 2019.

[209] M. Ye and Y. Guo. Zero-shot classification with discriminative semantic representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7140–7148, 2017.

[210] S. W. Yoon, J. Seo, and J. Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. *arXiv preprint arXiv:1905.06549*, 2019.

[211] A. W. Yu, D. Dohan, Q. Le, T. Luong, R. Zhao, and K. Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2, 2018.

[212] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 534–544, 2019.

[213] Y. Yuan, X. Lan, L. Chen, W. Liu, X. Wang, and W. Zhu. A closer look at temporal

sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021.

[214] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10287–10296, 2020.

[215] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6688–6697, 2020.

[216] D. Zhang, X. Dai, and Y.-F. Wang. Metal: Minimum effort temporal activity localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3882–3892, 2020.

[217] H. Zhang, A. Sun, W. Jing, and J. T. Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.585`.

[218] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–666, 2016.

[219] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–11, 2017.

[220] S. Zhang, J. Su, and J. Luo. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1230–1238, 2019.

[221] S. Zhang, H. Peng, J. Fu, and J. Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12870–12877, 2020.

[222] Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2021.

[223] Z. Zhang, Z. Lin, Z. Zhao, J. Zhu, and X. He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4098–4106, 2020.

[224] Z. Zhang, Z. Zhao, Z. Lin, X. He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[225] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015.

[226] J. Zhao, D. Lu, K. Ma, Y. Zhang, and Y. Zheng. Deep image clustering with category-style representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[227] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017.

[228] Y. Zhao, Z. Zhao, Z. Zhang, and Z. Lin. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206, 2021.

[229] H. Zhou, C. Zhang, Y. Luo, Y. Chen, and C. Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8454, 2021.

[230] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.