Deep Learning with Insufficient Data

YUTONG CAO

Supervisor: Prof. Dacheng Tao
Auxiliary Supervisor: Dr. Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

This research reported in this thesis was supported by
the award of a Research Training Program scholarship
to the PhD Candidate.

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

July 2024

Statement of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.
I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Yutong Cao

Authorship Attribution Statement

This thesis contains several chapters that were previously published by Yutong Cao (Yu-Tong Cao) in the following publications:

(1) Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16 (pp. 230-247). Springer International Publishing* Presented in Chapter 3. I designed the research, developed the code to implement the methods, analyzed the data, conducted the experiments, and wrote the draft of the paper.

(2) Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-IID data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22279-22289).* Presented in Chapter 5. I designed the research, developed the code to implement the methods, analyzed the data, conducted the experiments, and wrote the draft of the paper.

This thesis also contains a chapter that is currently under review and made public online:

(1) Yu-Tong Cao, Jingya Wang, Baosheng Yu, and Dacheng Tao. Responsible Active Learning via Human-in-the-loop Peer Study. *arXiv preprint arXiv:2211.13587 (2022).* Presented in Chapter 4. I designed the research, developed the code to implement the methods, analyzed the data, conducted the experiments, and wrote the draft of the paper.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Yutong Cao

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Dacheng Tao

Abstract

The recent success of deep learning approaches mainly relies on the availability of huge computational resources and large-scale labelled data. However, the persistent challenge of data insufficiency remains a significant concern within the deep learning research community. This thesis delves into multiple facets of data insufficiency across various learning tasks.

Without sufficient data for training, deep models cannot fully explore the data space. Due to the cost or infeasibility of collecting training data, many real-world applications face the problem of insufficient training data. In this thesis, data insufficiency is explored and tackled within the attribute-based person search task in Chapter 3. In scenarios where individuals are described based on attributes such as gender, age, clothing, or accessories in surveillance data, the goal is to retrieve individuals that match specific attribute descriptions. A significant obstacle arises from the non-overlapping nature of most training and testing classes (attribute combinations). It is impossible to come across all possible attribute combinations during training since the training data cannot cover the whole space of possible attribute combinations. To address this data insufficiency, we present a novel data augmentation technique that generates synthetic features for unseen attribute combinations.

The insufficiency of data labels also leads to challenges in the learning of deep models. The subsequent chapters primarily explore solutions to this issue. While unannotated data may be relatively accessible for some tasks, data labelling is often prohibitively expensive. To mitigate this issue, active learning emerges as a central theme in later chapters of this thesis. Active learning involves the strategic selection of the most informative labelled data for annotation, aiming to optimize the utility of the annotation budget.

Among the many perspectives of active learning studies, one that lacks attention is privacy-aware active learning. Many applications where highly confidential personal data or medical data are used could benefit from active learning, since the privacy requirements usually increase the annotation costs. Proposing active learning methods that effectively solve the

label insufficiency problem without breaching the privacy requirements is a good problem to address. Chapter 4 introduces a responsible active learning method that incorporates considerations for data privacy protection in scenarios involving cloud services. We can strictly protect a subset of local data while still utilizing it to update the active sampler without violating data privacy. To further ensure all local data remains untouched, Chapter 5 studies the federated active learning problem where local clients learn in a decentralized way without sharing local data, conventional active learning methods become ineffective due to non-IID data distribution. A method that tries to actively label local data for the global benefit is proposed in the chapter. The introduced method addresses the problem with a novel knowledge-specialized reweighting.

The insufficiency of data labels also occurs when the data requires complicated annotations that demand more well-trained professionals, which can be costly. An example would be human-object interaction (HOI) detection, which involves identifying human-object pairs and discerning interaction types within images featuring multiple interactions and objects across diverse scenes. Such annotation work cannot be accomplished by any random person. In Chapter 6, the insufficiency of data labels in HOI detection is addressed using active learning. What is difficult for the humans can also be difficult for deep models. Challenges primarily arise from the need to assess image informativeness based on various components such as object labels, bounding boxes, and interaction labels, as well as the presence of spurious features within images that may hinder the active sampling process. In this chapter, a method that comprehensively measures data informativeness is introduced.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dacheng Tao, and my auxiliary supervisor, Dr. Baosheng Yu, for their invaluable guidance, unwavering support, and encouragement throughout the duration of this candidature. Their expertise, patience, and mentorship have been instrumental in shaping my research journey and academic growth.

I am deeply thankful to my collaborator, Dr. Jingya Wang, for her insightful feedback, constructive criticism, and valuable contributions to the works. her expertise and dedication have enriched the quality of this thesis.

I am also grateful for the financial support from the Research Training Program. This funding has enabled me to pursue my research and achieve the goals outlined in this thesis.

I extend my appreciation to my friends and collaborators for their encouragement, support, and camaraderie throughout this endeavour. Their friendship and camaraderie have provided a source of motivation and inspiration during challenging times.

Finally, I would like to express my deepest gratitude to my family for their unwavering love, encouragement, and understanding. Their unwavering support and belief in my abilities have been a constant source of strength and motivation.

This thesis would not have been possible without the support and contributions of all those mentioned above. I am truly grateful for their kindness, guidance, and encouragement.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Motivation

The recent success witnessed in the field of deep learning has largely been attributed to the abundance of both computational resources and large-scale data. However, despite the remarkable progress achieved, the challenge of data insufficiency persists as a significant concern within the deep learning research community.

Data insufficiency in deep learning generally falls into two categories: lack of sufficient training data and lack of sufficient data labels.

Firstly, there may not be enough data available for training, which hampers the model's ability to adequately cover all possible scenarios during testing. This limitation can lead to models that struggle to generalize effectively to unseen data, hindering their real-world applicability and performance.

One of the deep learning applications that suffer from insufficient training data would be person search. The person search problem usually uses descriptive queries to retrieve person images from a large pool. The major task is to find the best match between the semantic queries and the person images. Ideally, the model should learn from all sorts of person images to develop a better representation space for matching. However, this is almost impossible in reality due to the numerous possibilities of appearances and the limited training data. Aside from collecting more and more data, a more achievable solution is to make the model 'imagine' unseen appearances based on existing knowledge, just like how human beings do.

In later chapters, we show how this is conducted with a nature-inspired method that allows 'imagining' and retrieving at the same time.

The second type of data insufficiency in deep learning is the shortage of data labels. Even when a sufficient amount of raw data is available, there may be an insufficiency of data annotations. In such cases, the data lacks the necessary labels or annotations required to train deep learning models effectively. This deficiency can severely limit the model's ability to learn meaningful patterns and relationships within the data, ultimately impacting its performance and reliability. One of the effective solutions is semi-supervised learning, where unlabelled data are trained to exploit everything available in the training data. However, these methods may highly rely on the quantity and quality of the existing data labels. They possibly introduce noise when handling the unlabelled data. Semi-supervised learning may not be the most suitable solution to insufficient data labels in certain circumstances. Another popular solution to the shortage of data annotations is active learning, which is also what is intensively studied in this thesis. Active learning, as the name suggests, actively samples unlabelled data for annotation to ensure the annotation budget is efficiently utilized. Despite the simple aim, it is never easy to determine what kind of unlabelled data is more valuable when their labels are obtained. This becomes a even more difficult problem when active learning needs to be conducted in more complicated settings with realistic concerns.

- **How to protect data privacy in active learning?** It is a common practice to pass all unlabelled data to a task model for active sampling score computation. This results in potential data leakage risk when the task model utilizes external resources or needs to be shared with external parties. Active learning using cloud resources utilizes external resources. A relevant example would be the photo management systems on smartphones, where users save their photos locally but wish to have them correctly sorted as they like. Active learning could provide a cheap and accurate solution to this problem. External hardware resources are usually important since the hardware on smartphones can be limited. However, there may exist personal photos that the users would never want to share. In this case, to propose a method that simultaneously prevents sending all data to an external resource provider while still

maintaining competitive active sampling ability is essential. Another case is where the models need to be shared across different parties, such as federated learning. In federated learning, many local clients maintain their own local datasets. Only the model parameters are shared with a central server for aggregation to achieve a more powerful model. Federated active learning is a very practical setting. A closely related application is learning with hospital data. Hospital data is usually difficult to be labelled and highly confidential. Federated active learning would be a very effective way to solve the task. In different hospitals, it is common that the data distributions differ. It is important to find a federated active learning method which not only protects local data privacy, but also effectively samples informative unlabelled data that are informative to the aggregated model.

- **How to handle applications with complex labels in active learning?** For tasks like classification or segmentation, the informativeness of data is relatively easy to quantify with certain metrics like entropy scores. As human beings, it is also possible to tell that some images can be classified into multiple categories, making them likely to be informative for classification. However, there are also tasks that annotate images with complicated labels, such as Human-Object Interaction (HOI) detection, for example. The task requires annotating images with bounding boxes of objects and humans, the object categories, and the interaction categories. There may also be multiple interactions visible. Even for human beings, it is difficult to evaluate what kind of images are informative for this task. To tackle the active HOI detection problem, it is essential to find a comprehensive active sampling criterion.

Both of these challenges pose significant obstacles to the advancement of deep learning technologies and urgently require effective solutions. Addressing data insufficiency is crucial for unlocking the full potential of deep learning and enabling the development of more robust, reliable, and generalizable models capable of tackling real-world problems effectively.

1.2 Thesis Outline

The thesis offers several solutions to address data insufficiency in different learning scenarios and tasks. These solutions focus on two main aspects: (I) data augmentation to overcome data insufficiency, and (II) active learning to address data annotation insufficiency. Before getting into details, we provide background knowledge to help understand the concepts in this thesis in Chapter 2.

Part I: Data Augmentation. In Chapter 3, we introduce SAL which synthesizes data from unseen classes to deal with data insufficiency. In attribute-based person search which retrieves persons with attribute descriptions, a significant challenge arises from the non-overlapping nature of most training and testing classes (i.e., attribute combinations). During training, it is impossible to encounter all potential attribute combinations, as the training data cannot encompass the entirety of this diverse space. Consequently, this data insufficiency presents a hurdle to effective model training and performance.

Part II: Active Learning. Active learning aims to select the most informative unlabelled data for annotation. The problem is studied from three different aspects

- In Chapter 4, we propose a responsible active learning method that allows data privacy protection with cloud services in use. Our method allows for strictly protecting some of the unlabelled data without sacrificing the performance of the active sampler.
- In Chapter 5, a method for federated active learning, which allows decentralized active learning without violating local data privacy, is introduced. Our method addresses the non-IID data problem in federated active learning.
- In Chapter 6, an active learning method specifically designed for the Human-Object Interaction (HOI) detection task is proposed. We design a comprehensive active sampling method to solve the specific task.

We summarize and discuss the possible future studies in Chapter 6.

CHAPTER 2

Background

This chapter provides a comprehensive overview of the fundamental concepts and techniques used to address the data insufficiency problem in deep learning.

The first section introduces popular solutions for mitigating insufficient training data. This includes synthetic data generation and mixup, both enhance the quantity and diversity of the training dataset. Emphasis is placed on techniques that are particularly relevant to the themes and methods discussed in this thesis, demonstrating their application and efficacy in various scenarios.

The second section introduces active learning, one of the major strategies for overcoming the insufficiency of training labels. Active learning involves selectively querying the most informative data points for labeling, thereby maximizing the efficiency of the labeling process. This section will cover different active learning strategies, such as uncertainty sampling, query-by-committee, and diversity-based approaches, highlighting their advantages and challenges. Additionally, the integration of active learning with other data insufficiency solutions will be explored, showcasing how these methods can work with each other to enhance model performance.

2.1 Insufficiency of Training Data

With insufficient training data, the most straightforward solution is to augment the training dataset. In the famous fundamental work AlexNet [82], data augmentation techniques like cropping, flipping, and colour editting are adopted. Similar techniques are widely used in

later works [57, 58]. Although these simple techniques generally improve performance, they may not be powerful enough in complicated tasks. In this section, the techniques to augment training data by synthesizing and mixing up data will be introduced.

2.1.1 Synthetic Data

To learn a model with limited training data, using synthesized data as augmentation is an intuitive solution. It is widely adopted in one-shot[31], few-shot[55, 173, 44], and zero-shot[107, 84, 207, 36] learning tasks. In this thesis, the insufficiency of training data is mainly studied on the attribute-based person search problem, to which the zero-shot learning task is the most related with the attribute-annotated data and the encountering of unseen categories. Wang et al. [173] proposed a meta-learning based end-to-end method that generates new images and learns the classifier simultaneously. However, the generated images are not forced to be discriminative. Thus can be similar with the training data. Kumar et al. [84] introduced a encoder-decoder model that can generate new images given attributes. It does not discriminate whether the images are real or fake. In this case, the quality of the synthetic images is not controlled. Overall, synthetic images may suffer from distortion, and it is time-consuming to train a good generator for images. Therefore, many works synthesize data in feature level. Dixit et al. [31] proposed the attribute-guided augmentation method which benefits the training of a one-shot recognition model by synthesizing features of given attribute values in the feature space. However, it requires training an extra attribute regressor and only two attributes of multiple classes are used in the experiment. Hariharan et al. [55] introduced generating new examples by transferring modes of variations using analogies in base classes under a few-shot setting. The performance is not as good as the later GAN-based models, possibly because the collected analogies are not reliable. Gao et al. [44], proposed a GAN-based method which synthesizes novel class features as a mixture of related base classes and reached state of the art for the few-shot learning task. However, the few-shot methods cannot utilise any semantic data like in the cross-modal retrieval problems due to the property of the task. Utilising semantic attributes effectively, Long et al. [107] proposed a method that directly maps the semantic features to the visual feature space for synthesizing in

zero-shot learning. However, it is a one-to-one mapping that limits the synthetic data size and provides no variations for the synthetic data. Felix et al. [36] proposed a GAN-based model, but with a cycle-consistency constraint that forces the synthetic visual features to be generated back to the original semantic features. The final classifier is trained with the synthetic visual features separately. This method only synthesizes the visual features and embeds the features in the visual space.

2.1.2 Mixup

Mixup [193] is proven effective in generating virtual features for augmentation in supervised learning originally. It encourages the model to perform convex behaviors on training examples. Later, mixup has been extended and applied in semi-supervised learning [160, 7, 123, 6], domain adaptation [112, 137], open-set learning [201], and cross-domain person re-identification [109]. Berthelot et al. [7] proposed MixMatch to handle semi-supervised learning. MixMatch adopts mixup to mix both labelled data unlabelled data for augmentation. ReMixMatch [6] improved MixMatch with distribution alignment and augmentation anchoring. RealMix [123] aims at improving semi-supervised learning in more realistic settings. However, these semi-supervised learning methods together with MixUp only deal with the closed-set classification task that involve only data from a fixed set of classes. Luo et al. [109] proposed the cross-domain mixup scheme to interpolate the source and target domains as an intermediate state. Although, this method deals with a retrieval problem, it still does not handle cross-modality data. Most of these follow-up works of Mixup use the technique to match the distributions of either labelled and unlabelled data or data from multiple domains.

2.2 Active Learning

Active learning has been widely adopted to address the insufficiency of training labels in deep learning. By focusing on selecting the most informative unlabeled data for annotation, active learning effectively controls annotation costs within budget constraints, making it a practical

solution for many machine learning applications such as image classification [97, 133, 99], object detection [117], and semantic segmentation [145, 11, 155].

Existing active learning methods mainly fall in three categories, uncertainty-based [42, 185, 148, 192, 5, 128, 170, 75, 34], disagreement-based [141, 39, 22, 116, 27, 26], and diversity-based [140, 3]. Aside from the conventional active learning methods that are usually evaluated with the image classification task, there are also active learning methods that are specifically designed for other computer vision applications such as semantic segmentation [145, 180, 125], tracking [191], and object detection [20, 190, 175, 117]. Siddiqui et al. [145] proposed to consider active learning for semantic segmentation in a multi-view setting by measuring the inconsistency of predictions across views as the model uncertainty. The work of Yuan et al. [191] introduced a novel active learning method for object tracking in videos. The method integrates the temporal relationship of the moving object in the video sequence for active sampling. To consider both the localization and the classification information in active object detection, Choi et al. [20] combined the aleatoric and epistemic uncertainties. The former measures the intrinsic noise in images and the latter computes the lack of knowledge in the model. Active teacher [117] proposed by Mi et al. dealt with the object detection problem via pseudo-labelling using an exponential mean average teacher. The sampling metric includes both the uncertainty of the box prediction and that of the object classification.

CHAPTER 3

Symbiotic Adversarial Learning for Attribute-based Person Search

---

In this chapter, we study the attribute-based person search problem, a complex and data-insufficient task that merits thorough investigation in this thesis. Attribute-based person search is in significant demand for applications where no detected query images are available, such as identifying a criminal from witnesses. However, the task itself is quite challenging because there is a huge modality gap between images and physical descriptions of attributes. Often, there may also be a large number of unseen categories (attribute combinations). The available training data can only cover a small proportion of the possible categories. The current state-of-the-art methods either focus on learning better cross-modal embeddings by mining only seen data, or they explicitly use generative adversarial networks (GANs) to synthesize unseen features. The former tends to produce poor embeddings due to insufficient data, while the latter does not preserve intra-class compactness during generation. In this chapter, we present a symbiotic adversarial learning framework, SAL. Two GANs sit at the base of the framework in a symbiotic learning scheme: one synthesizes features of unseen classes/categories, while the other optimizes the embedding and performs the cross-modal alignment on the common embedding space. Specifically, two different types of generative adversarial networks learn collaboratively throughout the training process and the interactions between the two mutually benefit each other. This mutual beneficial relationship between the two networks imitates the symbiotic relationship in the nature, e.g., the relationship between the crocodile and plover birds, where the latter is fed on the food residue of the former, and the former gets teeth cleaned by the latter. Our two networks in SAL also form a relationship where the networks benefit from each other. To further improve the symbiotic adversarial

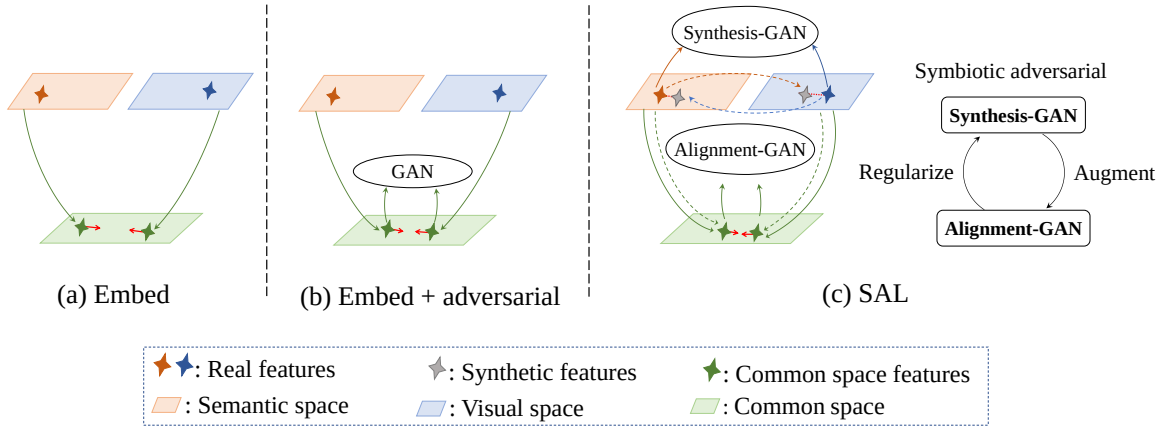


FIGURE 3.1: Three categories of cross-modal retrieval with common space learning: (a) The embedding model that projects data from different modalities into common feature space. (b) The embedding model with common space adversarial alignment. (c) The proposed SAL jointly considers feature-level data augmentation and cross-modal alignment by two GANs.

learning process, we propose a more advanced SAL+ framework. We introduce Symbiotic Adversarial Mixup (SAM) which utilizes the symbiosis nature of the SAL framework to augment the learning and form SAL+. SAM helps to enrich the learning space in SAL by producing reliable paired data. Extensive evaluations show our proposed networks' superiority over nine state-of-the-art methods with two challenging pedestrian benchmarks, PETA and Market-1501.

3.1 Introduction

The goal of "person search" is to find the same person in non-overlapping camera views at different locations. In surveillance analysis, it is a crucial tool for public safety. To date, the most common approach to person search has been to take one detected image captured from a surveillance camera and use it as a query [95, 164, 194, 59, 200, 151, 96, 18]. However, this is not realistic in many real-world applications – for example, where a human witness has identified the criminal, but no image is available.

Attributes, such as gender, age, clothing or accessories, are more natural to us as searchable descriptions, and these can be used as soft biometric traits to search for in surveillance

data [104, 87, 69, 134]. Compared to the queries used in image-based person search, these attributes are also much easier to obtain. Further, semantic descriptions are more robust than low-level visual representations, where changes in viewpoint or diverse camera conditions can be problematic. Recently, a few studies have explored sentence-based person search [96, 93]. Although this approach provides rich descriptions, unstructured text tends to introduce noise and redundancy during modeling. Attributes descriptions, on the other hand, are much cheaper to collect, and they inherently have a robust and relatively independent ability to discriminate between persons. As such, attribute descriptions have the advantage of efficiency over sentence descriptions in person search tasks.

Unfortunately, applying a cross-modal, attribute-based person search to real-world surveillance images is very challenging for several reasons. (1) There is a huge semantic gap between visual and textual modalities (i.e., attribute descriptions). Attribute descriptions are of lower dimensionality than visual data, e.g., tens versus thousands, and they are very sparse. Hence, in terms of data capacity, they are largely inferior to visual data. From performance comparisons between single-modal retrieval (image-based person search) and cross-modal retrieval (attribute-based person search) using current state-of-the-art methods, there is still a significant gap between the two, e.g., mAP 84.0% [138] vs. 24.3% [32] on Market-1501 [198, 102]. (2) Most of the training and testing classes (attribute combinations) are non-overlapping, which leads to zero-shot retrieval problems – a very challenging scenario to deal with [32]. Given an attribute-style query, model aims to have more capacity to search an unseen class given a query of attributes. (3) Compared to the general zero-shot learning settings one might see in classification tasks, surveillance data typically has large intra-class variations and inter-class similarities with only a small number of available samples per class (Fig. 3.5), e.g., ~ 7 samples per class in PETA [30] and ~ 26 samples per class in Market-1501 compared with ~ 609 samples per class in AWA2 [178]. One category (i.e., general attribute combinations that are not linked to a specific person) may include huge variations in appearance – just consider how many dark-haired, brown-eyed people you know and how different each looks. Also, the inter-class distance between visual representations from different categories can be quite small considering fine-grained attributes typically become ambiguous in low resolution with motion blur.

One general idea for addressing this problem is to use cross-modal matching to discover a common embedding space for the two modalities. Most existing methods focus on representation learning, where the goal is to find projections of data from different modalities and combine them in a common feature space. This way, the similarity between the two can be calculated directly [54, 167, 37] (Fig. 3.1(a)). However, these approaches typically have a weaker modality-invariant learning ability and, therefore, weaker performance in cross-modal retrieval. More recently, some progress has been made with the development of GANs [50]. GANs can better align the distributions of representations across modalities in a common space [163], plus they have been successfully applied to attribute-based person search [184]. (Fig. 3.1(b)). There are still some bridges to cross, however. With only a few samples, only applying cross-modal alignment to a high-level common space may mean the model fails to capture variances in the feature space. Plus, the cross-modal alignment probably will not work for unseen classes. Surveillance data has all these characteristics, so all of these problems must be overcome.

To deal with unseen classes, some recent studies on zero-shot learning have explicitly used GAN-based models to synthesize those classes [107, 84, 207, 36]. Compared to zero-shot classification problems, our task is cross-modal retrieval, which requires learning a more complex search space with a finer granularity. As our experiments taught us, direct generation without any common space condition may reduce the intra-class compactness.

Hence, we present a fully-integrated symbiotic adversarial learning framework for cross-modal person search, called SAL. Inspired by symbiotic evolution, where two different organisms cooperate so both can specialize, we jointly explore the feature space synthesis and common space alignment with separate GANs in an integrated framework at different scales (Fig. 3.1(c)). The proposed SAL mainly consists of two GANs with interaction: (1) A synthesis-GAN that generates synthetic features from semantic representations to visual representations, and vice versa. The features are conditioned on common embedding information so as to preserve very fine levels of granularity. (2) An alignment-GAN that optimizes the embeddings and performs cross-modal alignment on the common embedding space. In this way, one GAN augments the data with synthetic middle-level features, while

the other uses those features to optimize the embedding framework. Meanwhile, a new regularization term, called common space granularity-consistency loss, forces the cross-modal generated representations to be consistent with their original representations in the high-level common space. These more reliable embeddings further boost the quality of the synthetic representations. To address zero-shot learning problems when there is no visual training data for an unseen category, we use new categories of combined attributes to synthesize visual representations for augmentation. An illustration of feature space data augmentation with a synthesized unseen class is shown in Fig. 3.2.

To further deal with the aforementioned challenges in attribute-based person search, this work additionally proposes the Symbiotic Adversarial Mixup (SAM) which produces mixed features to fit in the SAL framework for feature space augmentation utilizing the symbiosis nature. This technique is designed to generate paired mixed features that lie out of the training data neighbourhood in feature space to facilitate the cross-modal retrieval task. The symbiosis nature of SAL constraints this feature generation process, in order to produce features that benefit our task learning. We aim to enrich the sparse feature space and to close the semantic gap between modalities by providing extra paired mixed data in SAL for the attribute-based person search task. An illustration of feature space data augmentation with SAM is shown in Fig. 3.2. we extensively evaluate SAM in the SAL framework with different widths and depths. The work also proposes supervised cross-modal contrastive loss for a superior common space embedding base to improve the learning.

In summary, the work in this chapter makes the following main **contributions** to the literature:

- We propose a fully-integrated framework called symbiotic adversarial learning (SAL) for attribute-based person search. The framework jointly considers feature-level data augmentation and cross-modal alignment by two GANs at different scales. Plus, it handles cross-model matching, few-shot learning, and zero-shot learning problems in one unified approach.
- We introduce a symbiotic learning scheme where two different types of GANs learn collaboratively and specially throughout the training process. By generating qualified data to optimise the embeddings, SAL can better align the distributions in

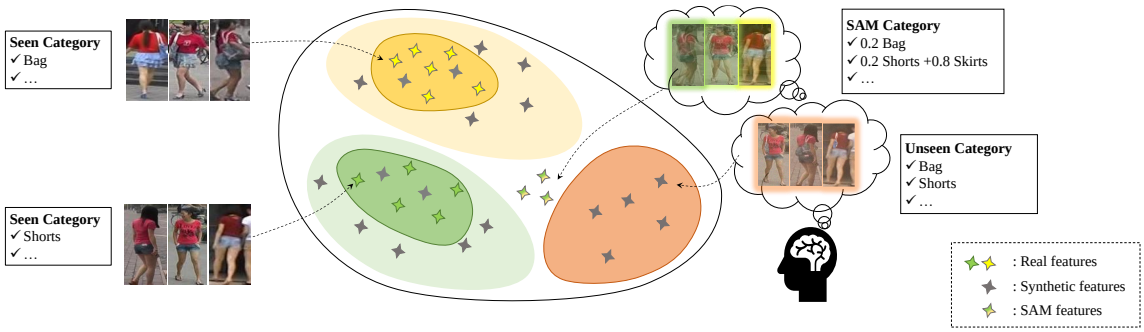


FIGURE 3.2: An illustration of feature space data augmentation with a synthesized unseen class and a mixed class from Symbiotic Adversarial Mixup (SAM).

the common embedding space, which, in turn, yields superior cross-modal search results.

- We introduce a Symbiotic Adversarial Mixup to further explore the out of the training data neighbourhood. With our Symbiotic Adversarial Mixup, we are able to accommodate attribute combinations and images that are quite different from the training data.
- Extensive evaluations on the PETA [30] and Market-1501 [198] datasets demonstrate the superiority of SAL at attribute-based person search over nine state-of-the-art attribute-based models.

3.2 Related work

Attribute-Based Person Search. In recent years, semantic attribute recognition of pedestrians has drawn increasing attention [90, 85, 29, 165, 196] and it has been extensively exploited for image-based person search as a middle-level feature [86, 149, 150, 102, 166]. Some studies exploit attribute-based person search for cross-modal retrieval [159, 146, 139, 184, 32, 71]. Early attribute-based retrieval methods intuitively rely on attribute prediction. For example, Vaquero et al. [159] proposed a human parts based attribute recognition method for person retrieval. Siddiquie et al. [146] utilized a structured prediction framework to integrate ranking and retrieval with complex queries, while Scheirer et al. [139] constructed normalized "multi-attribute spaces" from raw classifier outputs. However, the pedestrian

attribute recognition problem is far from being solved [90, 165, 105]. Consistently predicting all the attributes is a difficult task with sparsely-labeled training data: the images of people from surveillance cameras are low resolution and often contain motion blur. Also the same pedestrian would rarely be captured in the same pose. These "imperfect attributes" are significantly reducing the reliability of existing models. Shi et al. [144] suggested transfer learning as a way to overcome the limited availability of costly labeled surveillance data. They chose fashion images with richly labeled captions as the source domain to bridge the gap between unlabeled pedestrian surveillance images. They were able to produce semantic representations for a person search without annotated surveillance data, but the retrieval results were not as good as the supervised/semi-supervised methods, which raises a question of cost versus performance. [184] posed attribute-based person search as a cross-modality matching problem. They applied an adversarial leaning method to align the cross-modal distributions in a common space. However, their model design did not extend to the unseen class problem or few-shot learning challenges. Dong et al. [32] formulated a hierarchical matching model that fuses the similarity between global category-level embeddings and local attribute-level embeddings. Although they consider a zero-shot learning paradigm in the approach, the main disadvantage of this method is that it lacks the ability to synthesize visual representations of an unseen category. Therefore, it does not successfully handle unseen class problem for cross-modal matching. More recently, Jeong et al. [71] proposed to reduce the modality gap between images and attributes using deep metric learning. The method helps to better align the image and attributes pairs, but does not put enough attention on exploring unobserved data that cannot be covered by the training set.

Cross-Modal Retrieval. Our task of searching for people using attribute descriptions is closely related to studies on cross-modal retrieval as both problems require the attribute descriptions to be aligned with image data. This is a particularly relevant task to text-image embedding [181, 168, 96, 163, 199, 156], where canonical correlation analysis (CCA) [54] is a common solution for aligning two modalities by maximizing their correlation. With the rapid development of deep neural networks (DNN), a deep canonical correlation analysis (DCCA) model has since been proposed based on the same insight [2]. Yan et al. [181] have subsequently extended the idea to an image-text matching method based on DCCA.

Beyond these core works, a variety of cross-modal retrieval methods have been proposed for different ways of learning a common space. Most use category-level (categories) labels to learn common representations [184, 163, 207, 199, 199]. However, what might be fine-grained attribute representations often lose granularity, and semantic categories are all treated as being the same distance apart. Several more recent works are based on using a GAN [50] for cross-modal alignment in the common subspace [163, 184, 207, 157]. The idea is intuitive since GANs have been proved to be powerful tools for distribution alignment [157, 61]. Further, they have produced some impressive results in image translation [206], domain adaptation [157, 61] and so on. However, when only applied to a common space, conventional adversarial learning may fail to model data variance where there are only a few samples and, again, the model design ignores the zero-shot learning challenge.

Mixup. Mixup [193] and its follow-up works [160, 7, 123, 6, 112, 137, 201, 109] has been introduced in Chapter 2. Most of these works use the technique to match the distributions of either labelled and unlabelled data or data from multiple domains. Our Symbiotic Adversarial Mixup is designed to accommodate the cross-modal retrieval problem by enriching the cross-modal feature spaces. It helps the learning of the middle-level and high-level GANs by producing paired augmentations in both modalities while utilizing the symbiosis nature of SAL.

3.3 Preliminaries

Problem Definition Our deep model for attribute-based person search is based on the following specifications. There are N labeled training images $\{\boldsymbol{x}_i, \boldsymbol{a}_i, y_i\}_{i=1}^{N}$ available in the training set. Each image-level attribute description $\boldsymbol{a}_i = [a_{(i,1)}, \ldots, a_{(i,n_{\text{attr}})}]$ is a binary vector with n_{attr} number of attributes, where 0 and 1 indicate the absence and presence of the corresponding attribute. Images of people with the same attribute description are assigned to a unique category so as to derive a category-level label, specifically $y_i \in \{1, ..., M\}$ for M categories. The aim is to find matching pedestrian images from the gallery set $\mathcal{X}_{\text{gallery}} = \{\boldsymbol{x}_j\}_{j=1}^{G}$ with G images given an attribute description $\boldsymbol{a}_q$ from the query set.

3.4 Symbiotic adversarial learning

3.4.1 Multi-modal Common Space Embedding Base

One general solution for cross-modal retrieval problems is to learn a common joint subspace where samples of different modalities can be directly compared to each other. As mentioned in our literature review, most current approaches use category-level labels (categories) to generate common representations [184, 163, 207]. However, these representations never manage to preserve the full granularity of finely-nuanced attributes. Plus, all semantic categories are treated as being the same distance apart when, in reality, the distances between different semantic categories can vary considerably depending on the similarity of their attribute descriptions.

Thus, we propose a common space learning method that jointly considers the global category and the local fine-grained attributes. The common space embedding loss is defined as:

$$\mathcal{L}_{\text{embed}} = \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{att}}. \tag{3.1}$$

The category loss utilizing the Softmax Cross-Entropy loss function for image embedding branch is defined as:

$$\mathcal{L}_{\text{cat}} = -\sum_{i=1}^{N} \log\Big(p_{\text{cat}}(\boldsymbol{x}_i, y_i)\Big), \tag{3.2}$$

where $p_{\text{cat}}(\boldsymbol{x}_i, y_i)$ specifies the predicted probability on the ground truth category y_i of the training image $\boldsymbol{x}_i$.

To make the common space more discriminative for fine-trained attributes, we use the Sigmoid Cross-Entropy loss function which considers all m attribute classes:

$$\begin{aligned}\mathcal{L}_{\text{att}} = -\sum_{i=1}^{N}\sum_{j=1}^{m}\Big(& a_{(i,j)} \log\left(p_{\text{att}}^{(j)}(\boldsymbol{x}_i)\right) \\ & + (1 - a_{(i,j)}) \log\left(1 - p_{\text{att}}^{(j)}(\boldsymbol{x}_i)\right)\Big),\end{aligned} \tag{3.3}$$

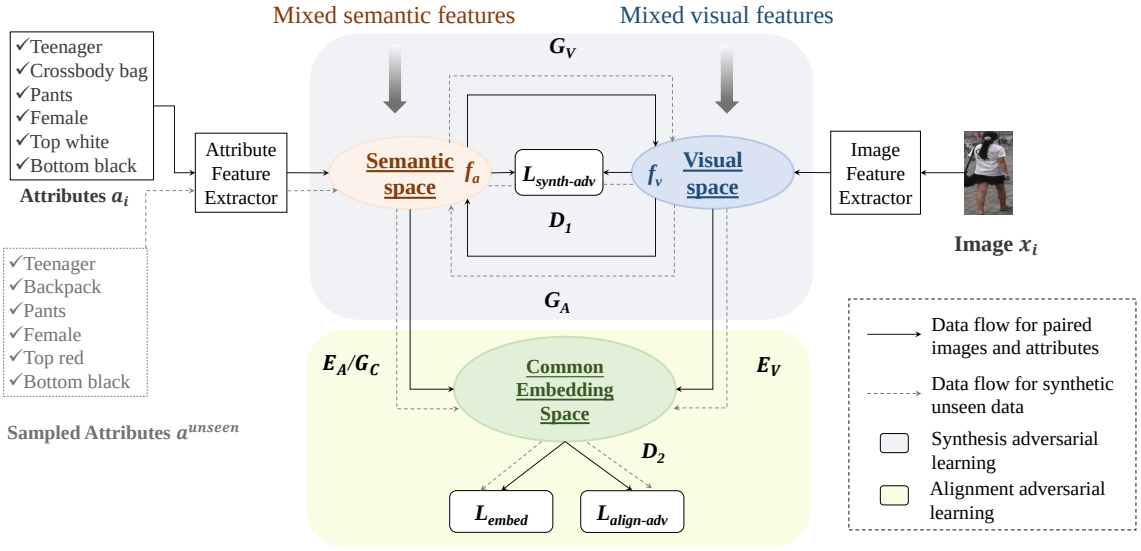


FIGURE 3.3: An overview of the proposed SAL+ Framework. Symbiotic Adversarial Mixup is fit into the SAL framework.

where $a_{(i,j)}$ and $p_{\text{att}}^{(j)}(\boldsymbol{x}_i)$ define the ground truth label and the predicted classification probability on the j-th attribute class of the training image $\boldsymbol{x}_i$, i.e. $\boldsymbol{a}_i = [a_{(i,1)}, \cdots, a_{(i,m)}]$ and $\boldsymbol{p}_{\text{att}}^{(j)} = [p_{\text{att}}^{(1)}(\boldsymbol{x}_i), \cdots, p_{\text{att}}^{(m)}(\boldsymbol{x}_i)]$. Similar loss functions of Eq. (3.2) and Eq. (3.3) are applied to the attribute embedding branch as well.

The multi-modal common space embedding can be seen as our base, named as *Embed* (Fig. 3.1(a)). A detailed comparison is shown in Table 3.2.

To further improve the common space embedding, we propose a supervised cross-modal contrastive loss which is developed from the supervised contrastive loss [73].

$$\mathcal{L}_{\text{cross-sup}}(\boldsymbol{r}) = \log \frac{\exp(\boldsymbol{r} \cdot \hat{\boldsymbol{r}}_p / \tau)}{\sum_{b \in B} \exp(\boldsymbol{r} \cdot \hat{\boldsymbol{r}}_b / \tau)}, \tag{3.4}$$

where $\boldsymbol{r}$ represents normalized semantic features in the common embedding space and $\hat{\boldsymbol{r}}$ represents normalized visual features. Specifically, $\hat{\boldsymbol{r}}_p$ represents positive visual features that belong to the same category as $\boldsymbol{r}$ within the batch B, including the visual feature that pairs with $\boldsymbol{r}$. And $\hat{\boldsymbol{r}}_b$ represents all visual features in batch B. The temperature $\tau \in \mathcal{R}^+$ is a hyperparameter. By updating the multi-modal common space embedding using $\mathcal{L}_{\text{embed+}} = \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{att}} + \mathcal{L}_{\text{cross-sup}}$, we can obtain an improved base, named as *Embed+*.

3.4.2 Middle-Level Granularity-Consistent Cycle Generation

To address the challenge of insufficient data and bridge the modality gap between middle-level visual and semantic representations, we aim to generate synthetic features from the semantic representations to supplement the visual representations and vice versa. The synthetic features are conditioned on common embedding information so as to preserve very fine levels of granularity for retrieval. The middle-level representations are denoted as f_v for visual and f_a for attribute. More detail is provided in Fig. 3.3.

As paired supervision are available for seen categories during generation, conventional unsupervised cross-domain generation methods, e.g. CycleGAN [206], DiscoGAN [77] are not perfect suit. To better match joint distributions, we consider single discriminator distinguishes whether the paired data (f_a, f_v)) is from a real feature distribution $p(f_a, f_v)$ or not. This is inspired by Triple Generative Adversarial Networks [21]. The TripleGAN works on semi-supervised classification task with a generator, a discriminator and a classifier while our synthesis-GAN consists of two generators and a discriminator. As shown in Fig. 3.3, our synthesis-GAN consists of three main components: Generator G_A synthesizes semantic representations from visual representations; Generator G_V synthesizes visual representation from semantic representations, and a discriminator D_1 focused on identifying synthetic data pairs. Given that semantic representations are rather more sparse than visual representations, we add a noise vector $z \sim \mathcal{N}(0, 1)$ sampled from a Gaussian distribution for G_V to get variation in visual feature generation. Thus, G_V can be regarded as a one-to-many matching from a sparse semantic space to a rich visual space. This accords with the one-to-many relationship between attributes and images. The training process is formulated as a minmax game between three players - G_V, G_A and D_1 - where, from a game theory perspective, both generators can achieve their optima. The adversarial training scheme for middle-level cross-modal feature generation can, therefore, be formulated as:

$$\begin{aligned}\mathcal{L}^{\text{SAL}}_{\text{gan1}}(G_A, G_V, D_1) = &\mathbb{E}_{(f_a,f_v)\sim p(f_a,f_v))}[\log(D_1(f_a, f_v))] \\ &+ \frac{1}{2}\mathbb{E}_{f_a\sim p(f_a)}[\log(1 - D_1(f_a, \widetilde{f}_v))] \\ &+ \frac{1}{2}\mathbb{E}_{f_v\sim p(f_v)}[\log(1 - D_1(\widetilde{f}_a, f_v))].\end{aligned} \tag{3.5}$$

The discriminator D_1 is fed with three types of input pairs: (1) The fake input pairs $(\widetilde{f}_a, f_v)$, where $\widetilde{f}_a = G_A(f_v)$. (2) The fake input pairs $(f_a, \widetilde{f}_v)$, where $\widetilde{f}_v = G_V(f_a, z))$. (3) The real input pairs $(f_a, f_v) \sim p(f_a, f_v)$.

To constrain the generation process to only produce representations that are close to the real distribution, we assume the synthetic visual representation can be generated back to the original semantic representation, which is inspired by the CycleGAN structure [206]. Given this is a one-to-many matching problem as mentioned above, we flex the ordinary two-way cycle consistency loss to a one-way cycle consistency loss (from the semantic space to the visual space and back again):

$$\mathcal{L}^{\text{SAL}}_{\text{cyc}}(G_A, G_V) = \mathbb{E}_{f_a\sim p(f_a)}[||G_A(G_V(f_a, z)) - f_a||_2]. \tag{3.6}$$

Furthermore, feature generation is conditioned on embeddings in the high-level common space, with the aim of generating more meaningful representations that preserve as much granularity as possible. This constraint is a new regularization term that we call "the common space granularity-consistency loss", formulated as follows:

$$\begin{aligned}\mathcal{L}^{\text{SAL}}_{\text{consis}}(G_A, G_V) = &\mathbb{E}_{f_v\sim p(f_v)}[||E_A(\widetilde{f}_a) - E_V(f_v)||_2] \\ &+ \mathbb{E}_{f_a\sim p(f_a)}[||E_V(\widetilde{f}_v) - E_A(f_a)||_2] \\ &+ \mathbb{E}_{(f_a,f_v)\sim p(f_a,f_v)}[||E_A(\widetilde{f}_a) - E_A(f_a)||_2] \\ &+ \mathbb{E}_{(f_a,f_v)\sim p(f_a,f_v)}[||E_V(\widetilde{f}_v) - E_V(f_v)||_2],\end{aligned} \tag{3.7}$$

where E_A and E_V stand for the common space encoders for attribute features and visual features respectively.

Lastly, the full objective of synthesis-GAN is:

$$\mathcal{L}^{\text{SAL}}_{\text{synth-adv}} = \mathcal{L}^{\text{SAL}}_{\text{gan1}} + \mathcal{L}^{\text{SAL}}_{\text{cyc}} + \mathcal{L}^{\text{SAL}}_{\text{consis}}. \tag{3.8}$$

3.4.3 High-Level Common Space Alignment with Augmented Adversarial Learning

On the high-level common space which is optimized by Eq. (3.1), we introduce the second adversarial loss for cross-modal alignment, making the common space more modality-invariant. Here, E_A works as a synchronous common space encoder and generator. Thus, we can define E_A as G_C in this adversarial loss:

$$\begin{aligned}\mathcal{L}^{\text{SAL}}_{\text{gan2}}(G_C, D_2) =& \mathbb{E}_{f_v \sim p(f_v)}[\log D_2(E_V(f_v))] \\ &+ \mathbb{E}_{f_a \sim p(f_a)}[\log(1 - D_2(E_A(f_a)))].\end{aligned} \tag{3.9}$$

Benefit from the middle-level GAN that bridges the semantic and visual spaces, we can access the augmented data, $\widetilde{f}_a$ and $\widetilde{f}_v$, to optimise the common space, where $\widetilde{f}_a = G_A(f_v)$ and $\widetilde{f}_v = G_V(f_a, z)$. Thus the augmented adversarial loss is applied for cross-modal alignment by:

$$\begin{aligned}\mathcal{L}^{\text{SAL}}_{\text{aug1}}(G_C, D_2) =& \mathbb{E}_{f_a \sim p(f_a)}[\log D_2(E_V(\widetilde{f}_v))] \\ &+ \mathbb{E}_{f_v \sim p(f_v)}[\log(1 - D_2(E_A(\widetilde{f}_a))].\end{aligned} \tag{3.10}$$

And can be used to optimize the common space by:

$$\mathcal{L}^{\text{SAL}}_{\text{aug2}}(E_A, E_V) = \mathcal{L}_{\text{embed}}(\widetilde{f}_a) + \mathcal{L}_{\text{embed}}(\widetilde{f}_v). \tag{3.11}$$

The total augmented loss from the synthesis-GAN interaction is calculated by:

$$\mathcal{L}^{\text{SAL}}_{\text{aug}} = \mathcal{L}^{\text{SAL}}_{\text{aug1}} + \mathcal{L}^{\text{SAL}}_{\text{aug2}}. \tag{3.12}$$

The final augmented adversarial loss for alignment-GAN is defined as:

$$\mathcal{L}^{\text{SAL}}_{\text{align-adv}} = \mathcal{L}^{\text{SAL}}_{\text{gan2}} + \mathcal{L}^{\text{SAL}}_{\text{aug}}. \tag{3.13}$$

3.4.4 Symbiotic Training Scheme for SAL

Algorithm 1 summarizes the training process. As stated above, there are two GANs learn collaboratively and specially throughout the training process. The synthesis-GAN (including G_A, G_V and D_1) aims to build a bridge of semantic and visual representations in the middle-level feature space, and to synthesize features from both to each other. The granularity-consistency loss was further proposed to constrain the generation that conditions on high-level embedding information. Thus, the high-level alignment-GAN (including E_A, E_V and D_2) that optimizes the common embedding space benefits the synthesis-GAN with the better common space constrained by Eq. (3.7). Similarly, while the alignment-GAN attempts to optimize the embedding and perform the cross-modal alignment, the synthesis-GAN supports the alignment-GAN with more realistic and reliable augmented data pairs via Eqs. (3.10) and (3.11). Thus, the two GANs update iteratively with SAL's full objective becomes:

$$\mathcal{L}_{\text{SAL}} = \mathcal{L}^{\text{SAL}}_{\text{embed}} + \mathcal{L}^{\text{SAL}}_{\text{synth-adv}} + \mathcal{L}^{\text{SAL}}_{\text{align-adv}}. \tag{3.14}$$

Algorithm 1 Learning the SAL model.

Input: N labeled data pairs $(\boldsymbol{x}_i, \boldsymbol{a}_i)$ from the training set;
Output: SAL attribute-based person search model;

1: **for** $t = 1$ to *max-iteration* **do**
2: **Step 1: Multi-modal common space Embedding:** Updating both image branch and attribute branch by common space embedding loss $\mathcal{L}_{\text{embed}}$ (Eq. (3.1));
3: **Step 2: Middle-Level Granularity-Consistent Cycle Generation:** Updating G_A, G_V and D_1 by $\mathcal{L}_{\text{synth-adv}}$ (Eq. (3.8));
4: **Step 3: High-Level Common Space Alignment:** Updating E_A, E_V and D_2 by $\mathcal{L}_{\text{align-adv}}$ (Eq. (3.13));
5: **end for**

Semantic augmentation for unseen classes. To address the zero-shot problem, our idea is to synthesize visual representations of unseen classes from semantic representations. In contrast to conventional zero-shot settings with pre-defined category names, we rely on a myriad of attribute combinations instead. Further, this inspired us to also sample new attributes in the model design. Hence, during training, some new attribute combinations $\boldsymbol{a}^{unseen}$ are dynamically sampled in each iteration. The sampling for the binary attributes follows a Bernoulli distribution, and a 1-trial multinomial distribution for the multi-valued attributes (i.e., mutually-exclusive attributes) according to the training data probability. Once the attribute feature extraction is complete, f_a in the objective function is replaced with $[f_a, f_{a^{unseen}}]$, $f_{a^{unseen}}$ is used to generate synthetic visual feature via G_V. The synthetic feature pairs are then used for the final common space cross-modal learning.

3.5 Symbiotic Adversarial Mixup

The middle-level GAN augments the learning by generating extra features using training data features and using unseen attribute combinations. However, these generated features either lies near the training data or cannot provide enough pairing information for cross-modal alignment. To handle this, we further introduce Symbiotic Adversarial Mixup (SAM). Mixup [193] enriches the feature space using linearly combine the training examples. Such that the model can better predict examples that are not encountered in training data. However, Mixup is originally designed for single-modal classification problems, it does not consider

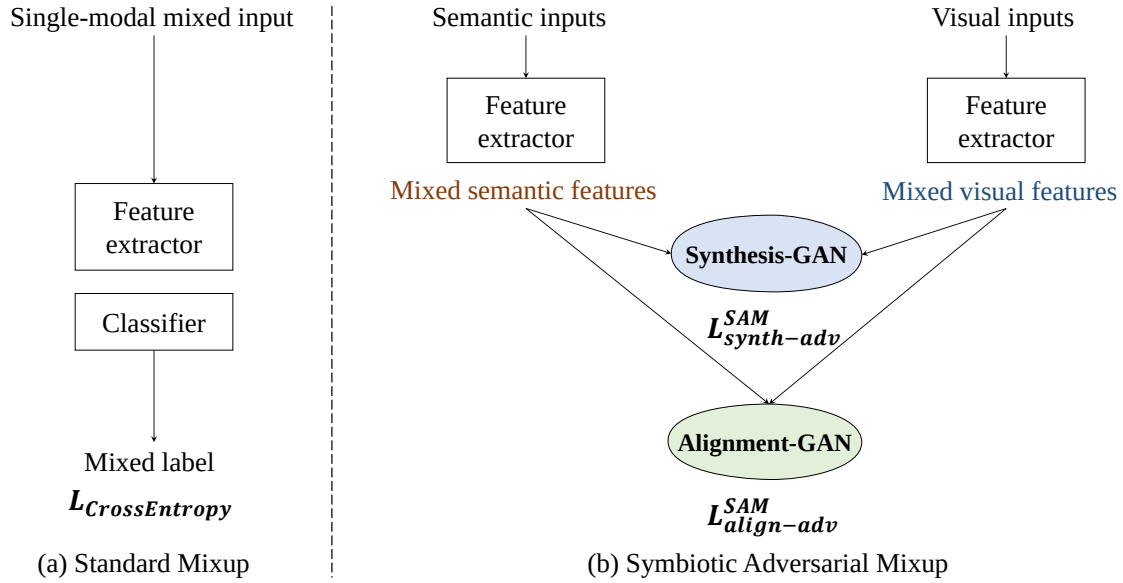


FIGURE 3.4: A comparison of standard Mixup which considers single-modal data and our Symbiotic Adversarial Mixup which takes cross-modal paired data.

multiple modalities. A possible solution is to apply Mixup on the attribute branch and the image branch respectively by using paired mixed data as input and apply classification loss with mixed labels as ground truths for each modality. Nevertheless, with the classification loss only, the pairing information may not be fully exploited for the retrieval task. We better utilize the symbiotic learning structure of SAL to close the semantic gap while benefiting from the mixed data. Hence, SAM is proposed. A comparison of previous Mixup and our Symbiotic Adversarial Mixup is illustrated in Fig. 3.4.

For SAM, we randomly sample a coefficient $\lambda \sim \mathrm{B}(\alpha, \beta)$ where $\mathrm{B}(\alpha, \beta)$ stands for the beta distribution with shape parameters α and β. We randomly sample two pairs of features. The first pair is denoted as visual feature f_{v_1} and semantic feature f_{a_1}, and the second is denoted as f_{v_2} and f_{a_2}. We denote the attribute annotation for the two pairs of features as $\mathbf{a}_1$ and $\mathbf{a}_2$. The mixed visual feature is then composed as $f_{vm} = \lambda f_{v_1} + (1-\lambda) f_{v_2}$, and the mixed semantic feature as $f_{am} = \lambda f_{a_1} + (1-\lambda) f_{a_2}$. The attribute annotation for f_{vm} and f_{am} is $\mathbf{a}_m = \lambda \mathbf{a}_1 + (1-\lambda)\mathbf{a}_2$. Here $\mathbf{a}_m$ is only theoretical and usually does not exist as a real attribute combination since the attributes are likely non-integers. The data flow for SAM is shown in Fig. 3.3. In the model design, we choose to mix data on feature-level rather than mixing them

on image-level (attribute-level) like in standard Mixup and many later mixup works. One reason is that [161] has proved manifold mixup of being able to produce smoother decision boundaries that are more distant to the training data. It means we can potentially explore further with feature-level mixup. The other reason to mix on feature-level is that mixing the attributes directly may lead to unreasonable attribute inputs. The original attributes signal the existence of certain appearances in an image with 0 or 1. Meanwhile, the numbers are binary and do not represent the intensity of each attribute property. Mixing them directly may produce decimal-valued attribute data that can not provide meaningful description for any possible pedestrian. Hence, we decide to have SAM mixed on feature-level.

We optimize with the paired mixed features by utilizing the middle-level cross-modal feature generation. The formulation is:

$$\begin{aligned}
&\mathcal{L}_{\text{gan1}}^{\text{SAM}}(G_A, G_V, D_1) \\
&= \mathbb{E}_{(f_{am}, f_{vm}) \sim p(f_{am}, f_{vm}))}[\log(D_1(f_{am}, f_{vm}))] \\
&+ \frac{1}{2}\mathbb{E}_{f_{am} \sim p(f_{am})}[\log(1 - D_1(f_{am}, \widetilde{f_{vm}}))] \\
&+ \frac{1}{2}\mathbb{E}_{f_{vm} \sim p(f_{vm})}[\log(1 - D_1(\widetilde{f_{am}}, f_{vm}))],
\end{aligned} \tag{3.15}$$

where $\widetilde{f_{vm}} = G_V(f_{am}, z)$ and $\widetilde{f_{am}} = G_A(f_{vm}, z)$. The discriminator D_1 then encounters three more types of input pairs related with the mixed features. Input pairs $(f_{am}, \widetilde{f_{vm}})$ and $(\widetilde{f_{am}}, f_{vm})$ are treated as fake pairs since features generated through the middle-level GAN are involved. Meanwhile, input pairs (f_{am}, f_{vm}) are treated as real pairs.

Similarly, we formulate the one-way cycle-consistency loss for SAM by enforcing the cycle-consistency behavior on mixed features as well:

$$\mathcal{L}_{\text{cyc}}^{\text{SAM}}(G_A, G_V) = \mathbb{E}_{f_{am} \sim p(f_{am})}[||G_A(G_V(f_{am}, z)) - f_{am}||_2]. \tag{3.16}$$

The common space granularity-consistency loss for SAM is formulated by enforcing the consistency of encoded mixed features in the common space. The formulation is:

$$\begin{aligned}
&\mathcal{L}_{\text{consis}}^{\text{SAM}}(G_A, G_V) \\
&= \mathbb{E}_{f_v \sim p(f_{vm})}[||E_A(\widetilde{f_{am}}) - E_V(f_{vm})||_2] \\
&+ \mathbb{E}_{f_{am} \sim p(f_{am})}[||E_V(\widetilde{f_{vm}}) - E_A(f_{am})||_2] \\
&+ \mathbb{E}_{(f_{am}, f_{vm}) \sim p(f_{am}, f_{vm})}[||E_A(\widetilde{f_{am}}) - E_A(f_{am})||_2] \\
&+ \mathbb{E}_{(f_{am}, f_{vm}) \sim p(f_{am}, f_{vm})}[||E_V(\widetilde{f_{vm}}) - E_V(f_{vm})||_2].
\end{aligned} \tag{3.17}$$

The full objective of synthesis-GAN for Symbiotic Adversarial Mixup is then formulated as:

$$\mathcal{L}_{\text{synth-adv}}^{\text{SAM}} = \mathcal{L}_{\text{gan1}}^{\text{SAM}} + \mathcal{L}_{\text{cyc}}^{\text{SAM}} + \mathcal{L}_{\text{consis}}^{\text{SAM}}. \tag{3.18}$$

The formulation of the adversarial loss in the high-level alignment-GAN in SAM is:

$$\begin{aligned}
\mathcal{L}_{\text{gan2}}^{\text{SAM}}(G_C, D_2) = \mathbb{E}_{f_{vm} \sim p(f_{vm})}[&\log D_2(E_V(f_{vm}))] \\
&+ \mathbb{E}_{f_{am} \sim p(f_{am})}[\log(1 - D_2(E_A(f_{am})))].
\end{aligned} \tag{3.19}$$

The mixed features also help to update the cross-modal alignment by using the SAM augmented adversarial loss:

$$\begin{aligned}
\mathcal{L}_{\text{aug1}}^{\text{SAM}}(G_C, D_2) = \mathbb{E}_{f_{am} \sim p(f_{am})}[&\log D_2(E_V(\widetilde{f_{vm}}))] \\
&+ \mathbb{E}_{f_{vm} \sim p(f_{vm})}[\log(1 - D_2(E_A(\widetilde{f_{am}})))].
\end{aligned} \tag{3.20}$$

Different from the features in the training set, the mixed features from Symbiotic Adversarial Mixup are not used for the prediction-level augmentation as in Eq. (3.11). This is different from the previous mixup methods, since they usually require the linear behavior in prediction space for mixed data. We choose not to apply this because our cross-modality task involves a semantic branch that predicts attributes and categories with attribute inputs. We believe that with such training, the classifier should be reliable enough. Further adding the embedding loss for mixed features may even harm the classifier. More of the challenges in this task fall in how to enrich the middle-level feature spaces and the common space. Hence, the total

augmented loss for alignment-GAN in SAM is formulated as:

$$\mathcal{L}_{\text{aug}}^{\text{SAM}} = \mathcal{L}_{\text{aug2}}^{\text{SAM}} + \mathcal{L}_{\text{aug1}}^{\text{SAM}}. \tag{3.21}$$

The final augmented adversarial loss for alignment-GAN in Symbiotic Adversarial Mixup is:

$$\mathcal{L}_{\text{align-adv}}^{\text{SAM}} = \mathcal{L}_{\text{gan2}}^{\text{SAM}} + \mathcal{L}_{\text{aug}}^{\text{SAM}}. \tag{3.22}$$

Finally, we name SAL with additional SAM to be SAL+ since SAM further augments the learning process of SAL. The full objective of SAL+ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{SAL+}} = & \mathcal{L}_{\text{embed}}^{\text{SAL}} \\ & + (1-\gamma)(\mathcal{L}_{\text{synth-adv}}^{\text{SAL}} + \mathcal{L}_{\text{align-adv}}^{\text{SAL}}) \\ & + \gamma(\mathcal{L}_{\text{synth-adv}}^{\text{SAM}} + \mathcal{L}_{\text{align-adv}}^{\text{SAM}}), \end{aligned} \tag{3.23}$$

where γ is the SAM coefficient that controls the amount of SAM loss in updating the full model.

3.6 Experiments

Datasets. To evaluate SAL with an attribute-based person search task, we used two widely adopted pedestrian attribute datasets[1]: (1) The ***Market-1501 attribute*** dataset [198] which the Market-1501 dataset annotated with 27 attributes for each identity [102] - see Fig. 3.5. The training set contains an average of ~ 26 labeled images in each category. During testing, 367 out of 529 **(69.4%)** were from **unseen** categories. (2) The ***PETA*** dataset [30] consists of $19,000$ pedestrian images collected from 10 small-scale datasets of people. Each image is annotated with 65 attributes. Following [184], we labeled the images with categories according to their attributes, then split the dataset into a training set and a gallery set by category. In the training set, there was on average ~ 7 labeled images in each category, and all 200 categories in the gallery set (**100%**) were **unseen**.

[1]The DukeMTMC dataset is not publicly available.

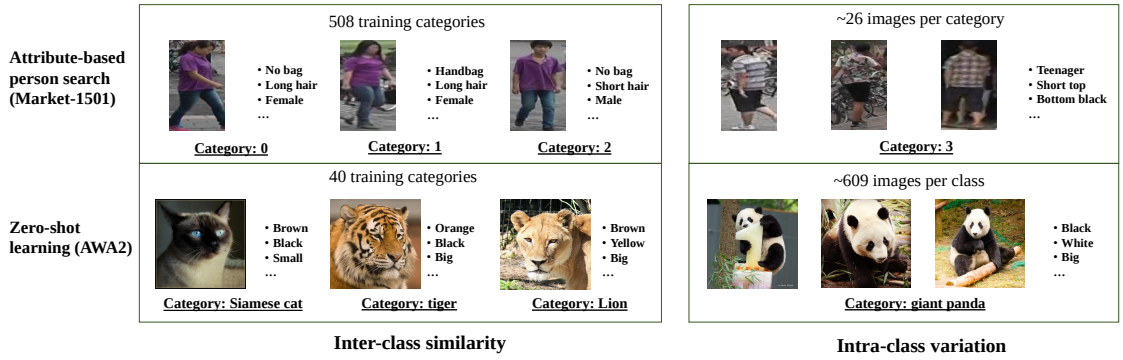


FIGURE 3.5: Example images from the person search Market-1501 [198] dataset and compared with the general zero-shot AWA2 [178] dataset.

TABLE 3.1: Attribute-based person search performance evaluation. Best results are shown in **bold**. The second-best results are underlined.

Metric (%)	Market-1501 Attributes				PETA			
Model	mAP	rank1	rank5	rank10	mAP	rank1	rank5	rank10
DeepCCA [2]	17.5	30.0	50.7	58.1	11.5	14.4	20.8	26.3
DeepMAR [90]	8.9	13.1	24.9	32.9	12.7	17.8	25.6	31.1
DeepCCAE [171]	9.7	8.1	24.0	34.6	14.5	14.2	22.1	30.0
2WayNet [35]	7.8	11.3	24.4	31.5	15.4	23.7	38.5	41.9
CMCE [93]	22.8	35.0	51.0	56.5	26.2	31.7	39.2	48.4
ReViSE [156]	17.7	24.2	45.2	57.6	31.1	30.5	57.0	61.5
MMCC [36]	22.2	34.9	58.7	70.2	33.9	33.5	57.0	69.0
AAIPR [184]	20.7	40.3	49.2	58.6	27.9	39.0	53.6	62.2
AIHM [32]	24.3	43.3	56.7	64.5	-	-	-	-
ASMR [71]	**30.1**	44.4	64.3	74.0	43.2	46.0	67.0	76.5
SAL [9]	29.8	49.0	68.6	**77.5**	41.2	47.0	66.5	74.0
SAL+	**30.1**	**51.0**	**69.6**	**77.5**	**48.8**	**58.5**	**77.5**	**84.5**

Performance Metric. The two metrics used to evaluate performance were the cumulative matching characteristic (CMC) and mean Average Precision (mAP). We followed [126] and computed the CMC on each rank k as the probe cumulative percentage of truth matches appearing at ranks $\leq k$. mAP measures the recall of multiple truth matches and was computed by first computing the area under the Precision-Recall curve for each probe, then calculating the mAP over all probes.

Implementation Details. SAL was implemented based on Torchreid [202, 204, 203] in the Pytorch framework [129] with ResNet-50 [57] as the image feature extractor. We used

fully connected (FC) layers to form the attribute feature extractor (64, 128, 256, 512), the middle-level generators (256, 128, 256, 512), and the encoders (512, 256, 128). Batch normalization and a ReLU nonlinear activation followed each of the first three layers with a Tanh (activation) before the output. The discriminators were also FC layers (512, 256, 1) with batch normalization and a leaky ReLU activation following each of the first two layers, and a Sigmoid activation prior to the output. After the common space embedding, the classifiers output with 1-FC layer. **Training process:** We first pre-trained the image branch and *Embed*/*Embed*+ model as a person search baseline (Fig. 3.4.1) until it converges. Then we trained the full SAL model for 60 epochs with a learning rate of 0.001 for the image branch and 0.01 for the attribute branch. We add SAM to SAL after 20 epochs of training. The mixing coefficient λ is sampled from the distribution $\mathrm{B}(2, 5)$ and the SAM coefficient γ is set to be 0.7. We chose Adam as the training optimizer [78] with a batch size of 128. During testing, we calculated the cosine similarity between the query attributes and the gallery images in the common embedding space with 128-D deep feature representations for matching. **Training time:** It took 77 minutes for SAL to converge (26.3M parameters) compared to 70 minutes for a single adversarial model Embed+adv to converge (25.0M parameters). Both training processes were run on the same platform with 2 NVIDIA 1080Ti GPUs.

3.6.1 Comparisons to the State-Of-The-Arts

The evaluation results of the attribute-based person search task appear in Table 3.1, presenting SAL's and SAL+'s performances in comparison to 10 state-of-the-art models across 4 types of methods. These are: (I) *attribute recognition* based method: (1) DeepMAR [90]. (II) *correlation* based methods: (2) Deep Canonical Correlation Analysis (DCCA) [2]; (3) Deep Canonically Correlated Autoencoders (DCCAE) [171]; (4) 2WayNet [35] (III) *common space embedding* : (5) Cross-modality Cross-entropy (CMCE) [93]; (6) ReViSE [156]; (7) Attribute-Image Hierarchical Matching (AIHM) [32]; (8) Adaptive Semantic Margin Regularizer (ASMR) [71] (note that the original paper's report uses different training and testing splits, hence we conduct the experiments with our splits for fair comparison) (IV) *adversarial*

learning: (9) Multi-modal Cycle-consistent (MMCC) model [36]; (10) Adversarial Attribute-Image Person Re-ID (AAIPR) [184].

From the results, we made the following observations: (1) SAL outperformed all 10 state-of-the-art models on both datasets in terms of rank1. On the Market-1501 dataset, the improvement over the second best scores were 4.6% (49.0 v 44.4). On the PETA dataset, the improvement was 1.0% (47.0 v 46.0). This illustrates SAL's overall performance advantages in cross-modal matching for attribute-based person search. (2) Directly predicting the attributes using the existing recognition models and matching the predictions with the queries is not efficient (e.g., DeepMAR). This may be due to the relatively low prediction dimensions and the sparsity problems with attributes in semantic space. (3) Compared to the common space learning-based method (CMCE), the conventional correlation methods (e.g., DCCA, DCCAE, 2WayNet) witnessed relatively poor results. This demonstrates the power of common space learning with a common embedding space. It is worth mentioning that CMCE is specifically designed to search for people using natural language descriptions, yet SAL outperformed CMCE by 7.0% mAP/14.0% rank1 with Market-1501, and by 15.0% mAP/15.3% rank1 with PETA. (4) The adversarial model comparisons, AAIPR and MMCC, did not fare well against SAL, especially AAIPR, which utilizes single adversarial learning to align the common space. This directly demonstrates the advantages of our approach with symbiotic adversarial design compared to traditional adversarial learning methods. (5) Among the compared state-of-the-arts, MMCC, ReViSE and AIHM addressed *zero-shot problems*. Against MMCC, SAL outperformed by 7.6% mAP/14.1% rank1 on Market-1501 and 7.3% mAP/13.5% rank1 on PETA. AIHM is the most recent state-of-the-art in this category of methods and SAL's performance improvement was 5.5% mAP/5.7% rank1 on Market-1501. This demonstrates the advantages of SAL's new regularization term, the common space granularity-consistency loss, for generating middle-level features. (6) Using deep metric learning, ASMR performs quite well by pulling embeddings from the same category together and pushing those from different categories away. SAL produced comparable results as ASMR since rank1 of SAL is better while mAP of ASMR is better. (7) With *Embed+* and SAM further added to SAL, the namely SAL+ achieved 2% and 11.5% improvement in rank1 in Market-1501 and PETA compared to SAL. It outperformed by 0.3% and 7.6% in rank1 for the two datasets

TABLE 3.2: Component analysis of SAL+ on PETA dataset.

Metric (%)	mAP	rank1	rank5	rank10
Embed	31.3	34.0	57.0	64.5
Embed + *adv*	35.0	37.5	60.5	66.5
Embed + *symb-adv*	40.6	44.0	64.0	70.5
Embed + *symb-adv* + *unseen*(SAL)	41.2	47.0	66.5	74.0
Embed+	43.7	47.5	68.5	77.0
SAL from *Embed+*	46.5	55.0	75.0	83.0
SAL+	48.8	58.5	77.5	84.5

respectively. The results show the superiority of introducing SAM which utilizes and benefits the symbiotic learning structure of SAL.

3.6.2 Further Analysis and Discussions

To further evaluate the different components in the model, we conduct studies on the PETA dataset.

Component analysis of SAL+. Here, we compared: (1) the base embedding model (*Embed*, Fig. 3.1(a)), which comprises the multi-modal common space embedding base (Sec. 3.4.1) and is optimized by embedding loss (Eq. (3.1)) only; (2) the base embedding model plus single adversarial learning (*Embed+adv*, Fig. 3.1(b)), (3) the base embedding model plus our symbiotic adversarial learning (*Embed+symb-adv*, Fig. 3.1(c)), and (4) our full SAL model, which includes the attribute sampling for unseen classes. The results are shown in Table 3.2.

Compared to the *Embed* model, *Embed+adv* saw an improvement of 3.7% mAP/3.5% rank1, whereas *Embed+symb-adv* achieved an improvement of 9.3% mAP/10% rank1, and SAL (*Embed+symb-adv+unseen*) witnessed a significant improvement of 9.9% mAP/13.0% rank1. This is a clear demonstration of the benefits of jointly considering middle-level feature augmentation and high-level cross-modal alignment instead of only having common space cross-modal alignment (41.2% vs 35.0% mAP and 47.0% vs 37.5% rank1). We also present the results of *Embed+*. On PETA, *Embed+* significantly outperformed the original base *Embed* with 12.4% and 13.5% improvement in mAP and rank1. Applying SAL based on *Embed+*, we further improve the results by 2.8% mAP/7.5% rank1. Finally, we add SAM

TABLE 3.3: Effect of interactions between two GANs on PETA dataset.

Metric (%)	mAP	rank1	rank5	rank10
SAL - $\mathcal{L}_{aug}$	35.4	38.0	60.0	69.0
SAL - $\mathcal{L}_{consis}$	35.2	39.5	56.5	66.0
SAL (Full interaction)	41.2	47.0	66.5	74.0

TABLE 3.4: Comparing stage-wise training vs. symbiotic training scheme.

Metric (%)	mAP	rank1	rank5	rank10
SAL w/ stage-wise training	35.0	41.0	58.0	65.0
SAL w/ symbiotic training	41.2	47.0	66.5	74.0

to achieve the SAL+ results and observe a further 2.3% mAP/3.5% rank1 improvement. This validates our idea that SAM can improve the learning by enriching the feature spaces at different levels for both modalities.

Effect of interactions between two GANs. Our next test was designed to assess the influence of the symbiotic interaction, i.e., where the two GANs iteratively regularize and augment each other. Hence, we removed the interaction loss between the two GANs, which is the total augmented loss from Eq. (3.12) and the common space granularity-consistency loss from Eq. (3.7). Removing the augmented loss (SAL - $\mathcal{L}_{aug}$) means the model no longer uses the augmented data. Removing the common space granularity-consistency loss (SAL - $\mathcal{L}_{consis}$) means the middle-level generators are not conditioned on high-level common embedding information, which should reduce the augmented data quality. The results in Table 3.3 show that, without augmented data, SAL's mAP decreased by 5.8% and by 6.0% without the granularity-consistency loss as a regularizer.

Comparing stage-wise training vs. symbiotic training scheme. We wanted to gain a better understanding of the power of the symbiotic training scheme (Sec. 3.4.4). So, we replaced the iterative symbiotic training scheme (SAL w/ symbiotic training) with a stage-wise training (SAL w/ stage-wise training). The stage-wise training [4] breaks down the learning process into sub-tasks that are completed stage-by-stage. So during implementation, we first train the synthesis-GAN conditioned on the common embedding. Then for the next stage, we optimise the alignment-GAN on the common space with synthetic data. As shown

TABLE 3.5: Effect of different mixing strategies.

Metric (%)	mAP	rank1	rank5	rank10
SAL from *Embed*+	46.5	55.0	75.0	83.0
+ Mixup	46.0	52.0	70.5	79.0
+ SAM (SAL+)	48.8	58.5	77.5	84.5

TABLE 3.6: Effect of mixing on different branches.

Metric (%)	mAP	rank1	rank5	rank10
SAL from *Embed*+	46.5	55.0	75.0	83.0
+ SAM image	48.2	56.0	74.0	83.0
+ SAM att	46.8	53.0	72.5	82.0
+ SAM (SAL+)	48.8	58.5	77.5	84.5

in Table 3.4, there was a 6.2% drop in mAP using the stage-wise training, which endorses the merit of the symbiotic training scheme. During the symbiotic training, the synthesized augmentation data and the common space alignment iteratively boost each other's learning. A better common space constrains the data synthesis to generate better synthetic data. And, with better synthetic data for augmentation, a better common space can be learned.

Effect of mixing strategies. To augment the learning with mixed features, there is also Mixup [193] aside from our Symbiotic Adversarial Mixup. We compare results from using the two strategies in Table 3.5. To add Mixup to SAL, we simply added mixed inputs separately to each branch in the model and applied classification with mixed label as ground truths. Results show that adding Mixup to SAL harms the learning. Adding Mixup reduces SAL's results (based on *Embed*+) by 0.5% mAP/3.0% rank1. This is possibly because SAL has already enriched the feature spaces with generated features and sampled attributes. Mixup may augment the learning with more data, but its simple design cannot further improve the common space learning. The results were outperformed by SAL+ by 2.8% mAP/6.5% rank1 on PETA dataset. SAM is more beneficial than Mixup when added to SAL possibly because SAM is designed to utilize the symbiotic learning structure and considers the paired property of mixed feature. Also, we show with experiment in later sections that SAM suits SAL better with the soft linearity enforcement instead of the hard linearity enforcement with classification loss.

TABLE 3.7: Effect of different mixing distributions.

Metric (%)	mAP	rank1	rank5	rank10
B$(2, 2)$	48.1	57.0	74.0	82.5
B$(2, 5)$	48.8	58.5	77.5	84.5
B$(2, 10)$	48.6	58.0	76.5	84.0

Effect of SAM on different branches. We also experimented by adding Symbiotic Adversarial Mixup to single branch in SAL and the results are presented in Table 3.6. Adding SAM to each branch of SAL separately does not affect the SAL results much. Adding SAM to SAL (based on *Embed*+) on image branch improved mAP and rank1 by 1.7% and 1.0%. Adding SAM on attribute branch increased the mAP by 0.3% and reduced the rank1 result by 2%. Maybe this is because the attribute branch is already well-learnt using SAL. The visual data can benefit more from mixed features at this stage due to the complexity of data. When SAM is added to both branches of SAL, allowing the mixed features to be fed in pairs, the results show a 2.3% improvement in mAP and a 3.5% improvement in rank1. SAM exploits the pairing information of the mixed features and closes the semantic gap. Therefore, SAL with whole SAM outperformed SAL with SAM on single branch with a clear margin.

Effect of different mixing distribution coefficients (α, β)**.** Since we sample the mixing coefficient λ in SAM from the distribution B(α, β), we are interested how the distribution parameters α and β can affect the results. The probability density function of Beta distribution has reflection symmetry such that $f(x; \alpha, \beta) = f(1 - x; \beta, \alpha)$, and the features to be mixed are randomly selected. Therefore, α and β can interchange without affecting the performance. We then fix α to 2 and tried three different values, 2, 5, and 10 for β. The distribution B$(2, 2)$ has a mean of 0.5, while B$(2, 5)$ has a mean of 0.28 and B$(2, 10)$ has a mean of 0.17. Results are shown in Table 3.7, all three outperform SAL. Using SAM constructed with λ sampled from B$(2, 5)$ produces the best result and using B$(2, 10)$ comes second.

Effect of different SAM coefficient γ**.** We also experimented using different values for the SAM coefficient γ. In Table 3.8, we show the results from using 0.1, 0.3, 0.5, 0.7, and 0.9 for the values of γ. Increasing the value of $gamma$ generally improves the performance when $gamma$ is less than 0.9. Using $\gamma = 0.7$ produced the best rank1 that is 3.0% better

TABLE 3.8: Effect of different SAM coefficient γ.

Metric (%)	mAP	rank1	rank5	rank10
$\gamma = 0.1$	47.8	55.5	74.0	82.5
$\gamma = 0.3$	48.0	56.0	74.5	82.5
$\gamma = 0.5$	48.3	57.0	74.0	83.5
$\gamma = 0.7$	48.8	58.5	77.5	84.5
$\gamma = 0.9$	49.2	57.5	75.5	84.0

TABLE 3.9: Effect of with and without common space embedding loss for SAM.

Metric (%)	mAP	rank1	rank5	rank10
with embedding loss	47.1	55.0	74.0	81.0
without embedding loss	48.8	58.5	77.5	84.5

than the rank1 result from $\gamma = 0.1$. The improvement in mAP is 1.0%. Using $gamma = 0.9$ produced the best mAP result which outperformed $gamma = 0.1$ by 1.4%. We eventually adopt $\gamma = 0.7$ since it generally performs the best. The experiments show that using SAM generally improves the performance, but a larger portion does not necessarily benefit the model more.

Effect of with and without common space embedding loss for SAM. We design our SAM to enforce the linearity of data in a soft manner, instead of applying the classification loss directly as in Mixup. We remove the embedding loss for SAM features. We show the results from SAM with and without embedding loss in Table 3.9. SAM with embedding loss means we are update the model with the common space embedding loss $\mathcal{L}_{embed}$ (eq. (3.1)) computed with the mixed features in the common space. With SAM, we are not able to compute the global category loss because the mixed features do not belong to any seen category. Then we have $\mathcal{L}_{embed} = \mathcal{L}_{att}$. This embedding loss is usually encouraged in the previous mixup works since they do not have the pairing information between cross-modal data to constrain the learning. However, in our case, we can see from the results that learning with the common space embedding loss for SAM does not improve the performance. This is probably because our target is to learn a better common embedding space for matching the images with the attributes instead of learning a better classifier. As long as we constrain the pairing relation between each pair of data for SAM, the exact attribute recognition ability may

be less prioritized. The embedding loss possibly hinders SAL from augmenting the feature space with SAM.

FIGURE 3.6: Ranked retrieval results. The query attributes are shown above the retrieved images. The green/red border represents correct/wrong selections respectively. The attributes in **bold** correspond to false matches.

Visualizing results. Fig. 3.6 visualizes the ranked results from *Embed*, *Embed+adv*, *Embed+symb-adv*, SAL, and SAL+ to qualitatively illustrate the performance of the three models. In the shown examples, SAL is able to pick out some correct images from candidates in the top 10 ranks, and SAL+ selected more with higher ranks. The others perform rather poorly, especially for the two examples on the right. Checking the wrong attributes of the

false ranks, we can see that SAL and SAL+ are more capable to discern fine-grained attributes, such as bags, hair, or gender. Moreover, we can tell from the top left example, SAL and SAL+ picked correct images of diverse appearances, which may indicate it has some ability to overcome intra-class variation problems. With mixed features from SAM, SAL+ has more enriched feature spaces at different levels. Possibly due to this reason, the rankings of correct images are higher from SAL+ compared to SAL. In the bottom right example in Fig. 3.6, *Embed*, *Embed+adv*, and *Embed+symb-adv* all selected quite a few male pedestrian images possibly because the attribute combination with 'short hair', 'top green' and other presented attributes are more often to be seen with 'male'. Our SAL and SAL+ perform better in identifying the gender attribute and selected a few correct pedestrian images in the top ten ranks.

3.7 Conclusion

In this work, we presented a novel symbiotic adversarial learning model, called SAL. SAL is an end-to-end framework for attribute-based person search. Two GANs sit at the base of the framework: one synthesis-GAN uses semantic representations to generate synthetic features for visual representations, and vice versa, while the other alignment-GAN optimizes the embeddings and performs cross-modal alignment on the common embedding space. Benefiting from the symbiotic adversarial structure, SAL is able to preserve finely-grained discriminative information and generate more reliable synthetic features to optimize the common embedding space. With the ability to synthesize visual representations of unseen classes, SAL is more robust to zero-shot retrieval scenarios, which are relatively common in real-world person search with diverse attribute descriptions. We further propose SAM which produces paired mixed features to update the synthesis-GAN and the alignment-GAN in the symbiotic learning structure. The symbiosis nature of the two GANs in SAL constraints the generation of extra paired features, allowing the generated mixed features to explore more of the feature space without sacrificing the quality. Features generated with SAM further enrich the sparse feature space and close the semantic gap between modalities in SAL for the attribute-based person search task. Extensive ablation studies illustrate the insights of our

model designs. Further, we demonstrate the performance advantages of SAL and SAL+ over a wide range of state-of-the-art methods on two challenging benchmarks.

CHAPTER 4

Responsible Active Learning via Human-in-the-loop Peer Study

In this chapter, an active learning method is proposed. As discussed in chatper 2, active learning allows efficient use of limited annotation budgets by selecting only informative unlabelled data for annotation. It effectively addresses the insufficiency of training labels with a controllable cost.

Recent active learning applications have benefited a lot from cloud computing services with not only sufficient computational resources but also crowdsourcing frameworks that include many humans in the active learning loop. However, previous active learning methods that always require passing large-scale unlabelled data to cloud may potentially raise significant data privacy issues. To mitigate such a risk, we propose a responsible active learning method, namely Peer Study Learning (PSL), to simultaneously preserve data privacy and improve model stability. Specifically, we first introduce a human-in-the-loop teacher-student architecture to isolate unlabelled data from the task learner (teacher) on the cloud-side by maintaining an active learner (student) on the client-side. During training, the task learner instructs the light-weight active learner which then provides feedback on the active sampling criterion. To further enhance the active learner via large-scale unlabelled data, we introduce multiple peer students into the active learner which is trained by a novel learning paradigm, including the In-Class Peer Study on labelled data and the Out-of-Class Peer Study on unlabelled data. Lastly, we devise a discrepancy-based active sampling criterion, Peer Study Feedback, that exploits the variability of peer students to select the most informative data to improve model stability. Extensive experiments demonstrate the superiority of the proposed

PSL over a wide range of active learning methods in both standard and sensitive protection settings.

4.1 Introduction

The recent success of deep learning approaches mainly relies on the availability of huge computational resources and large-scale labelled data. However, the high annotation cost has become one of the most significant challenges for applying deep learning approaches in many real-world applications. In response to this, active learning has been intensively explored by making the most of very limited annotation budgets. That is, instead of learning passively from a given set of data and labels, active learning first selects the most representative data samples that benefit a target task the most, and then annotates these selected data samples through human annotators, i.e., a human oracle. This active learning scheme has been widely used in many real-world applications, including image classification [97, 133, 99, 128] and semantic segmentation [155, 145, 11, 124].

In the past decades, artificial intelligence (AI) technology has been successfully applied in many aspects of daily life, while it also raises a new question about the best way to build trustworthy AI systems for social good. This problem is even more crucial to applications such as face recognition [127, 103, 143] and autonomous driving [68, 19, 12]. However, a responsible system has not received enough attention from the active learning community. Specifically, recent active learning applications have benefited a lot from cloud computing services with not only sufficient computational resources but also crowdsourcing frameworks that include many humans in the loop. The main issue is that existing methods require accessing massive unlabelled data for active sampling, which raises a significant data privacy risk, i.e., most local end users may feel worried and hesitate to upload large-scale personal data from client to cloud. Therefore, it is essential to develop a responsible active learning framework that is trustworthy for local end users.

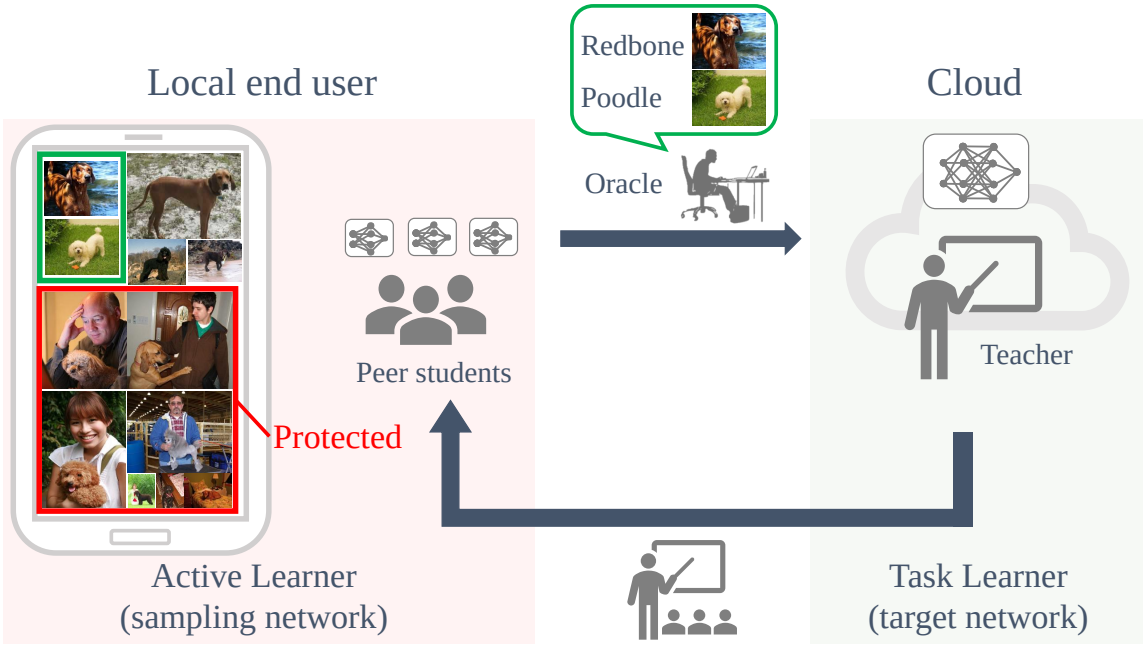


FIGURE 4.1: This framework empowers users to select a subset of data that will be strictly protected, accessible only to the local end user. This protected data can include sensitive user information such as facial images and home decor details. The selection of protected data is at the discretion of the users.

Since the main privacy risks derive from the active sampling process that requires accessing massive unlabelled data [5, 185, 3, 192, 128], we first introduce a human-in-the-loop teacher-student architecture for data isolation that separates the model into a task learner (a target network on the cloud-side) and an active learner (a sampling network on the client-side). An intuitive example is shown in Figure 4.1. At present, most active sampling solutions are based on either model uncertainty [42, 5, 185, 148, 192, 10] or data diversity [140, 3, 128] i.e., data with large uncertainty or data with features span over the space will be annotated by the oracle. Nevertheless, the above-mentioned methods usually rely on representations learned by a single model, potentially making the sampling prone to bias. Sampling actively with one model means that the result highly relies on quality of the single feature extractor. And when large models are unavailable due to limited resources, the active sampling results can be unsatisfying [23]. Inspired by disagreement-based active sampling approaches [39, 22, 116, 25, 26], we thus introduce multiple peer students to form a robust active learner as the sampling network. For active sampling, we consider a practical setting to give the maximum

protection to user data using personal album as a case study as follows: the local user has a large album with many personal photo images, where most sensitive photo images (e.g., 90% images) can be easily identified according to the tags such as timestamps and locations, and we call these images as the protected set. The objective of responsible active learning then is to select the most informative images from the remaining 10% images for annotation without uploading those 90% images to cloud. Therefore, the key to responsible active learning is to learn a proper active learner as the sampling network to train the task learner with very limited annotations.

To learn a good active learner, we introduce the proposed Peer Study Learning or PSL as follows. In real-world scenarios, all students not only learn with teacher guidance on course materials in classroom, but they also learn from each other after class using additional materials without teacher guidance. The difficulty of exercise questions in the additional materials can be estimated from the students' answers. When students cannot reach a consensus on the answer, the question is likely to be difficult and valuable. The students can provide the difficult exercise questions to the teacher as feedback. Inspired by this, the proposed Peer Study Learning consists of (1) In-Class Peer Study, (2) Out-of-Class Peer Study, and (3) Peer Study Feedback. Specifically, the In-Class Peer Study (Figure 4.2 (a)) allows peer students to learn from the teacher on labelled data. In Out-of-Class Peer Study (Figure 4.2 (b)), the peer students collaboratively learn in a group on unlabelled data without teacher guidance. The motivation is to enable the students' cooperation on easy questions while maintaining their dissent on challenging questions. After Peer Study, we use the discrepancy of peer students as Peer Study Feedback (Figure 4.2 (c)). All challenging data samples that cause a large discrepancy among peer students will be selected and annotated to improve the final model performance.

In this work, our main contributions can be summarized as follows:

- We introduce a new sensitive protection setting for responsible active learning.
- We propose a new framework PSL for responsible active learning, where a teacher-student architecture is applied to isolate unlabelled data from the task learner (teacher).

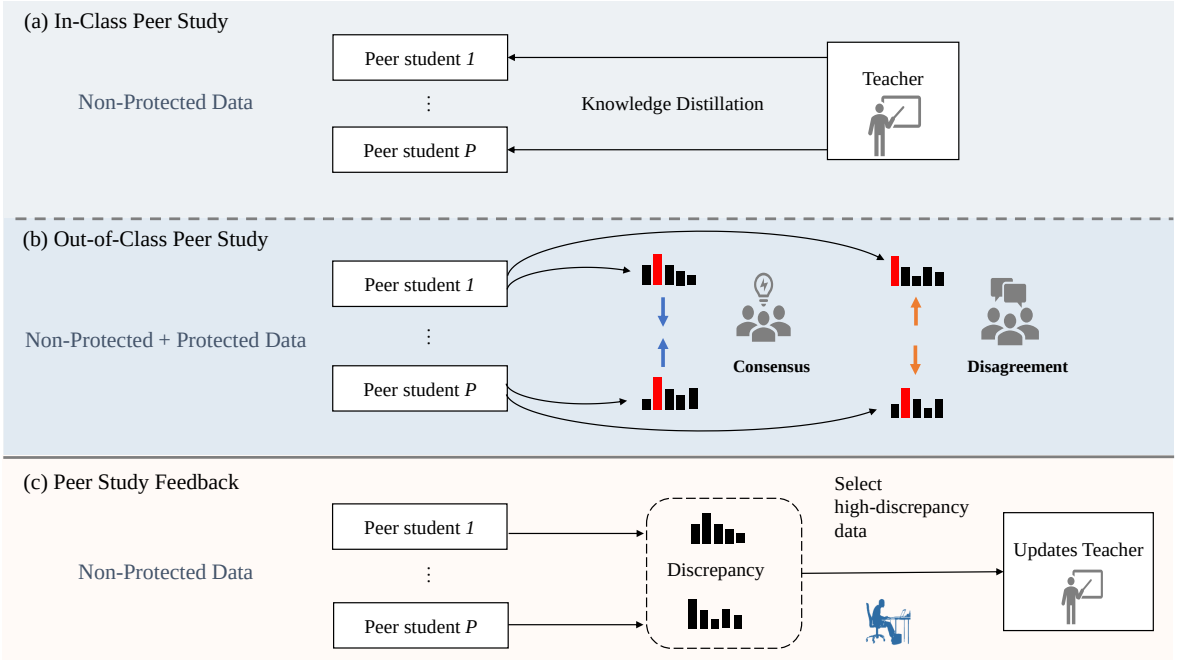


FIGURE 4.2: The main Peer Study Learning (PSL) pipeline with P peer students. For simplicity, the classification loss for the teacher and peer students are omitted. For Out-of-Class Peer Study, we provide a detailed illustration of the predictions with and without consensus on the prediction scores. The last row illustrates the active sampling process based on peer study feedback.

- We devise an effective learning strategy to mimic real-world in-class and out-of-class scenarios, i.e., In-Class Peer Study, Out-of-Class Peer Study, and Peer Study Feedback.
- We achieve state-of-the-art results with the proposed PSL in both standard and sensitive protection settings of active learning.

4.2 Related Work

Active Learning. Recent active learning methods can be broadly divided into three distinct categories: uncertainty-based [42, 5, 185, 148, 192, 170, 75, 10, 185, 148, 192, 170], diversity-based [140, 3, 128], and disagreement-based methods[39, 22, 116, 25, 26]. These methods were introduced in Chapter 2. Our Peer Study Learning falls in the category of disagreement-based methods. Previous disagreement-based active sampling methods are mainly applied on

classical machine learning models, such as decision tree and random forest, on small datasets, while directly applying these methods in deep neural networks result in a huge computational cost. Peer Study Learning uses a set of light-weight student models to build the active learner so as to reduce the cost, where we distil knowledge from the task learner to compensate possible performance drop.

Teacher-Student Architectures. A teacher-student architecture has been widely applied in knowledge distillation [60, 187, 154, 195, 119], transfer learning [188, 41, 182, 33, 179], and other tasks [8, 67, 153]. Specifically, [60] proposed a vanilla knowledge distillation method which uses a large pretrained teacher model to guide the learning of a small student model. To learn a good compressed student, [187, 188] proposed to distil knowledge from multiple teachers and train multiple students. [153] proposed Humble teacher where a teacher ensemble guides the student on strongly-augmented unlabelled data for semi-supervised object detection. Different from previous methods, our Peer Study Learning adopts a human-in-the-loop structure for active learning. A set of peer students benefit from not only teacher guidance via In-Class Peer Study but also the collaborative learning via Out-of-Class Peer Study using unlabelled data. The task learner (teacher) further benefits from the representative data samples selected by the active learner (student). Therefore, the teacher and peer students can be enhanced in such a responsible active learning loop.

4.3 Methods

In this section, we first provide an overview of responsible active learning. We then introduce different components of the proposed PSL framework in detail.

4.3.1 Overview

In active learning, we have a labelled data pool $\mathcal{D}_L = \{(x_i, y_i)\}_{i=0}^{N_L}$, and an unlabelled data pool $\mathcal{D}_U = \{x_i\}_{i=0}^{N_U}$. During learning, a batch of data $\mathcal{X} = \{x_i\}_{i=0}^{b}$ in $\mathcal{D}_U$ are actively sampled to be annotated by an oracle. The annotated data is then added to the labelled pool $\mathcal{D}_L$ and

removed from the unlabelled pool $\mathcal{D}_U$. The key to active learning is to identify the most informative data samples in $\mathcal{D}_U$ for annotation. In this work, we consider a more general setting to protect the user data as follows. For active sampling, we define a safe sampling pool $\mathcal{D}_S \subseteq \mathcal{D}_U$, i.e., all images belonging to the protected set $\mathcal{D}_U \setminus \mathcal{D}_S$ cannot be uploaded to cloud in any situation. Previous works usually adopt $\mathcal{D}_S = \mathcal{D}_U$, while we allow the local end users to decide the sampling pool $\mathcal{D}_S$ in a customized way such that it can be a small subset of D_U. To protect user data by isolating unlabelled data from cloud, we propose Peer Study Learning with an active learner on the client-side and a task learner on the cloud-side. The active learner consists of P peer students, and is parametrized with $\theta_p = \sum_{j=0}^{P} \theta_{p_j}$. the task learner is parametrized with θ_T. During training, θ_T can only be updated with labelled data, $\mathcal{D}_L$, while θ_p can be updated with both labelled and unlabelled data, $\mathcal{D}_L$ and $\mathcal{D}_U$. The main PSL pipeline is shown in Figure 4.2. In this task, we aim to obtain a powerful task learner with the assistance of cloud services for a local end-user that cannot access much training resources while keeping the annotation budget controllable. The large-scale task learner is placed on the cloud due to the insufficient local resources. Meanwhile, the light active learner locates locally to avoid violation of data privacy caused by uploading all unlabelled data to the cloud for active learning. As illustrated in Figure 4.1, the active learner samples unlabelled data locally and send them to the oracle for annotation. The annotated data are sent to the cloud to train the task learner.

4.3.2 In-Class Peer Study

In this subsection, we introduce the In-Class Peer Study process where the peer students learn from the teacher with labelled data. Since the task learner acts as both the target network and the teacher in Peer Study, we need to ensure it is well-learned. To train the task learner, we formulate the task learner's objective as:

$$\mathcal{L}_{\text{task}}(\theta_T; x_i) = \sum_{i=1}^{N} H(y_i, \sigma(f_{\theta_T}(x_i))) \tag{4.1}$$

where $(x_i, y_i) \in \mathcal{D}_L$ is the input data and label, σ is the softmax function, and f_{θ_T} stands for the task model. Function $H(\cdot)$ stands for cross-entropy loss, $H(p, q) = q \log(p)$ in the image-level classification task. In the pixel-level segmentation task, it represents the pixel-wise cross-entropy loss $H(p, q) = \sum_{\Omega} q \log(p)$, where Ω stands for all pixels.

In real-world classrooms, the teacher usually provides the students with questions and answers. The students can learn from the provided material with both the teacher's explanations and the solutions in the material. Analogously, we design to let our peer students learn from both the teacher's predictions and the data labels. We accordingly formulate the In-Class Peer Study loss as:

$$\begin{aligned} \mathcal{L}_{\text{in}}(\theta_p; x_i) = \sum_{j=1}^{P} \Big[& (1-\alpha) H\Big(\sigma(f_{\theta_{p_j}}(x_i)), y_i\Big) \\ & + \alpha\tau^2 H\Big(\sigma(\frac{f_{\theta_T}(x_i)}{\tau}), \sigma(\frac{f_{\theta_{p_j}}(x_i)}{\tau})\Big)\Big], \end{aligned} \tag{4.2}$$

where $(x_i, y_i) \in \mathcal{D}_L$ is the input data and label, the first term stands for the standard cross entropy classification loss using ground-truth labels and the second term stands for standard knowledge distillation loss using the task model outputs as the soft targets. α is a weight balancing the two terms. τ is the softmax temperature, larger τ implies softer predictions over classes. $f_{\theta_{p_j}}$ stands for a peer model with parameters θ_{p_j}.

4.3.3 Out-of-Class Peer Study

Motivated by that the students in real-world scenarios can learn from each other after class through a group study, we conduct Out-of-Class Peer Study using unlabelled data without teacher guidance. We define the set of consistent data if the predictions from any two peers reach a consensus as follows,

$$\begin{aligned} \mathcal{S}_A = \{ & x : \exists p_i, p_j \ s.t. \\ & \text{arg max}(f_{\theta_{p_i}}(x)) = \text{arg max}(f_{\theta_{p_j}}(x)), x \in \mathcal{D}_U\}. \end{aligned} \tag{4.3}$$

And for the set of inconsistent data that the predictions from all peers do not agree, we define it as follows,

$$\begin{aligned}\mathcal{S}_D =&\{x : \forall p_k, p_j \ s.t. \\ &\arg\max(f_{\theta_{p_k}}(x)) \neq \arg\max(f_{\theta_{p_j}}(x)), x \in \mathcal{D}_U\}.\end{aligned} \tag{4.4}$$

To adapt to pixel-level prediction tasks, we simply state that consensus is reached when more than half of the pixels have consistent predictions for any two peer students, and thus the datum belongs to consistent data. Otherwise, the datum belongs to inconsistent data. Then we formulate the Out-of-Class Peer Study loss as:

$$\begin{aligned}\mathcal{L}_{\text{out}}(\theta_p; x_A, x_D) = \ & max(0, -\mathbb{1}(d_D, d_A)(d_D - d_A) + \xi) \\ s.t. \ \mathbb{1}(d_D, d_A) = &\begin{cases} 1, & \text{if } d_D > d_A \\ -1, & \text{otherwise} \end{cases}\end{aligned} \tag{4.5}$$

where $d = \sum_{k \neq j}^{P} KL[\sigma(f_{\theta_{p_j}}(x)||\sigma(f_{\theta_{p_k}}(x)))]$ is the model discrepancy. d_D is the model discrepancy of a datum x_D randomly sampled from $\mathcal{S}_D$, and d_A is the model discrepancy of a datum x_A randomly sampled from $\mathcal{S}_A$. ξ is the margin for the Out-of-Class Peer Study loss. For pixel-level prediction, d is further summed over pixels of data. The Out-of-Class Peer Study loss re-ranks unlabelled data and pushes data with different label predictions on peers to have higher priority in active sampling. The simplified form of this loss is $\mathcal{L}_{\text{out}}(\theta_p; x_A, x_D) = \ max(0, -|d_D - d_A| + \xi)$. Combining this with our sampling criterion Peer Study Feedback, we are able to consider both the soft prediction scores and the hard prediction labels while conducting active sampling.

4.3.4 Peer Study Feedback

In real-life, even for students learning in the same class, the learning outcome for them can still vary a lot since they have different characters or different backgrounds. Therefore, in our In-Class Peer Study, despite learning from the same teacher, the P peer students' predictions may vary due to their independent initialization. And our Out-of-Class Peer Study further adjusts the prediction discrepancy among the students so that more challenging data can

Algorithm 2 Peer Study for Active Learning

Initialize: active learner θ_p, task learner θ_T
Hyperparameters: α, τ, ξ
Input: sample (X_L, Y_L) from labelled pool $\mathcal{D}_L$, (X_U) from unlabelled pool $\mathcal{D}_U$

for $e = 1$ to epochs **do**
  In-class Peer Study:
  Update the task learner with eq. (4.1) Update the active learner with eq. (4.2)
  Out-of-class Peer Study:
  Collect $\mathcal{S}_A$, $\mathcal{S}_D$ from X_U with (4.3) and eq. (4.4)
  Sample (x_A, x_D) from (S_A, S_D)
  Update active learner with eq. (4.5)
end for
return Trained θ_p and θ_T
run Peer Study Feedback to sample actively.

produce a larger discrepancy among students. Therefore, we can use the discrepancy among students to build our sampling strategy. The proposed method only uses the active learner to conduct the sample selection. Data from the unlabelled pool $\mathcal{D}_U$ are passed through our peer students and the discrepancy for any datum x_i among the peers is calculated according to the following selection criterion:

$$a(\theta_p, \mathcal{D}_U) = \sum_{k \neq j}^{P} KL[\sigma(f_{\theta_{p_j}}(x_i) || \sigma(f_{\theta_{p_k}}(x_i)))]. \tag{4.6}$$

The calculated discrepancies of data then serve as the sampling scores. We sort the sampling scores and select the top b data with largest sampling scores to send to the oracle for annotation and then add to the labelled pool $\mathcal{D}_L$. Note that in sensitive protection setting, we replace $\mathcal{D}_U$ with $\mathcal{D}_S$ in (4.6), only the non-protected data can be added to $\mathcal{D}_L$. We summarize the whole learning process for Peer Study in Alg. 2.

4.3.5 Discussion on Responsible Active Learning and Federated Learning

When it comes to privacy-preserving and responsible learning, federated learning has always been one of the solutions. In the federated learning scenario [115], all data are stored locally and metadata like model parameters or gradients can be transferred between the server and

the local device. However, transferring metadata can sometimes be risky in privacy protection. Model inversion attacks [38, 169], for example, can extract training data using these metadata. Thus, conventional federated learning approaches do not completely isolate specific local information from the server. Here, we consider more strict privacy preserving setting for responsible active learning. The proposed PSL keeps privacy data strictly untouched by the cloud side, not even through model parameters or metadata. Federated learning can avoid uploading raw data, while our PSL for responsible active learning can strictly protect certain extremely sensitive data with the cost of sharing a small amount of insensitive data with the task learner. Moreover, it is possible to extend our responsible approach with federated learning scenario. We show the design and results of this setting in Section 4.4.5.

4.4 Experiments

In this section, we first introduce a sensitive protection setting for responsible active learning. We then evaluate PSL in both sensitive protection and standard settings on image classification and semantic segmentation tasks. Lastly, we perform comprehensive ablation studies to better understand the proposed method.

4.4.1 Sensitive Protection Setting

We consider a sensitive protection setting in real-world scenarios as follows. A large portion of unlabelled data are protected by users to prevent from sending them to cloud. Data in the non-protected set can be selected and annotated by the oracle, then used to train the task learner. To simulate such a sensitive protection setting, we randomly select 90% data as the protected set, then the training starts with a labelled pool containing N_L images from the non-protected set. In each step, a batch of b data samples are selected by our active learner from the **unlabelled non-protected** set and sent to the oracle for annotation. The newly labelled data are then added to the labelled pool for next training steps. The above mentioned sampling process is terminated after the number of images in the labelled pool reaches a final N_f. A group of P peer students are used as the active learner to select data for annotation.

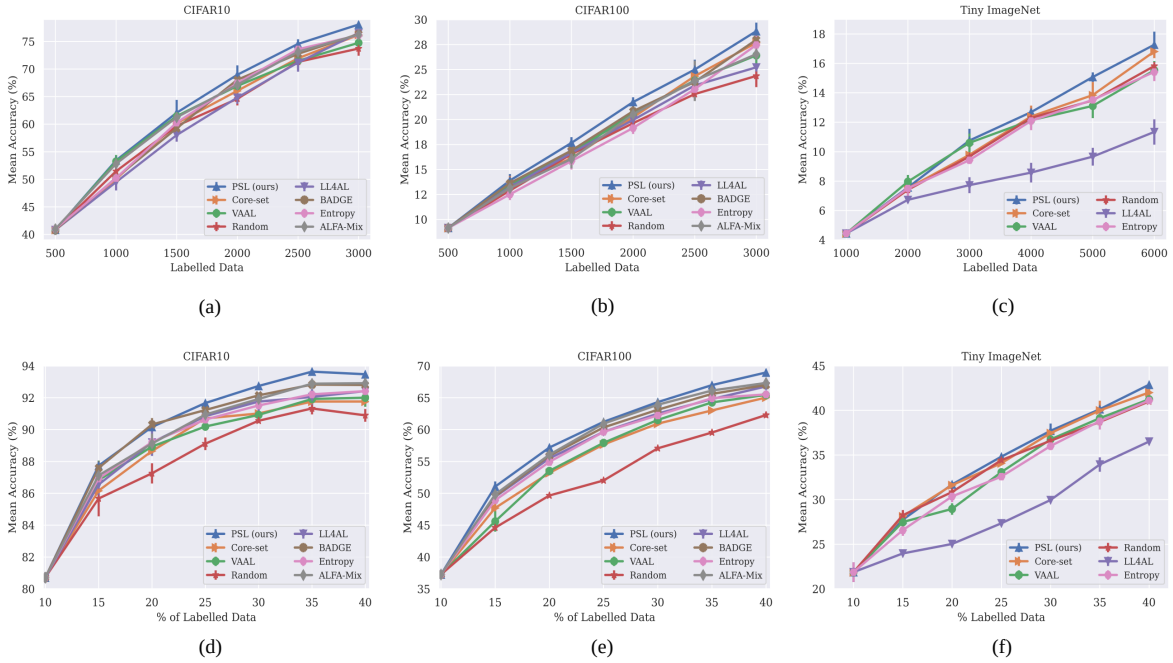


FIGURE 4.3: (a)-(c) Active learning in the sensitive protection setting on CIFAR10/100, and Tiny ImageNet. The x-axis shows the counts of labelled data. (d)-(f) Active learning in the standard setting on CIFAR10/100, and Tiny ImageNet. The x-axis shows the percentages of labelled data in all training data. The error bars represent the standard deviations among results from multiple different runs.

TABLE 4.1: A comparison of model complexity between PSL and others. 'CLS' and 'SEG' indicate the sampling models used for image classification and semantic segmentation, respectively.

Model	CLS	SEG
Entropy [142]	11.17M	-
BADGE [3]	11.17M	-
ALFA-Mix [128]	11.17M	-
Core-set [140]	11.17M	15.90M
LL4AL [185]	11.29M	16.02M
VAAL [148]	23.11M	23.11M
PSL (ours)	**0.16M**	**3.25M**

4.4.2 Implementation Details

For **image classification**, we evaluate PSL on CIFAR10/100 [80], containing $50,000$ training images and $10,000$ testing images from 10 and 100 different classes, respectively. The labelled pool is initialized with $N_L = 500$ and we sample $b = 500$ images at each step until $N_f = 3,000$. To evaluate PSL on data with more categories, we perform experiments on Tiny ImageNet [88] which contains $100,000$ images for training and $10,000$ images for evaluation from 200 different classes. The labelled pool is initialized with $N_L = 1,000$ and we sample $b = 1,000$ images at each step until $N_f = 6,000$. We use ResNet-18 [57] for the task learner and ResNet-8 [58] for the active learner. Accuracy is used as the evaluation metric. We train all models for 350 epochs with the weight $\alpha = 0.1$, softmax temperature $\tau = 4$, and the margin $\xi = 0.01$. For **semantic segmentation**, we evaluate PSL on Cityscapes [24], which contains $2,975$ images for training, and 500 images for testing. The initial labelled pool size is set as $N_L = 30$, the budget is $b = 30$, and the final labelled pool size is $N_f = 180$. We use DRN [189] for the task learner and MobileNetV3 [62] for the active learner. Mean IoU is used as the evaluation metric. The hyperparameters α, τ, and ξ are the same as those in image classification. We train all models for 50 epochs. If not otherwise stated, we use two peers in all our experiments and the reported results are averaged over five individual experimental runs. For fair comparison, we always use the same initial pools. We implement our method with Pytorch [129].

4.4.3 Active Learning in Sensitive Protection Setting

We compare PSL with (I) uncertainty-based methods, Entropy [142], LL4AL [185] and VAAL [148]; and (II) diversity-based methods, Core-set [140], BADGE [3] and ALFA-Mix [128]. We also report the performance of a simple random sampling baseline. We present the model complexity in Table 4.1, where our PSL has **the smallest model size** despite involving multiple peers. In addition, our PSL is also much more computationally efficient than other methods. We compare the sampling time that different methods use to sample the same amount of data from the same unlabelled pool in CIFAR10 [80] in the sensitive protection setting in Table 4.2. All experiments are conducted on the same NVIDIA GeForce

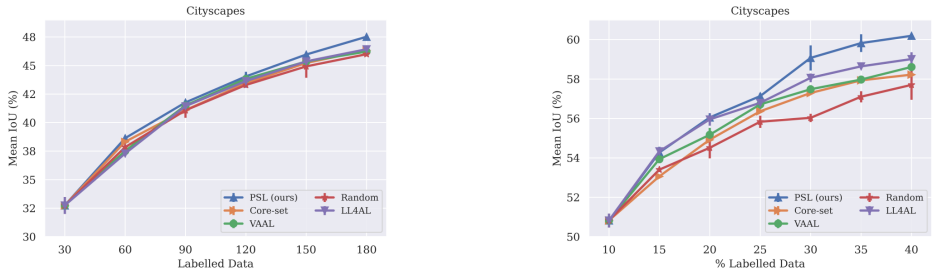
GTX 1080 Ti GPU for fair comparison. The results demonstrate that our PSL is much more efficient in sampling using the light-weight peer students as the active learner.

TABLE 4.2: Comparing the sampling time needed for different methods.

Model	Time (sec)
Entropy [142]	0.9
BADGE [3]	17.2
ALFA-Mix [128]	9.1
Core-set [140]	1.2
LL4AL [185]	8.9
VAAL [148]	2.5
PSL (ours)	0.001

Results on CIFAR10/100. As shown in Figure 4.3(a)-(b), our PSL outperforms all other methods on CIFAR10/100. On CIFAR10, all methods start with the initial labelled pool that produces the accuracy of 40.8%. With 1000 labelled data, PSL is still comparable with VAAL and ALFA-Mix. But when the labelled pool gets larger, PSL starts to outperform the rest methods with a visible margin. PSL achieves 78.06% with the eventual 3000 labelled data. The performance gap between PSL and the second best method LL4AL is around 1.5%. And there is a gap of 4.3% between PSL and the lowest results produced by random sampling. On CIFAR100, the initial labelled pool produces the accuracy of 9.1% for all methods. Similar with the situation on CIFAR10, PSL is close to other methods with 1000 labels and starts to show the superiority with more data annotated. PSL finally achieves 28.8% with 3000 labelled data. The performance gap between PSL and the second best method BADGE is around 0.9%. And there is a gap of 4.4% between PSL and the lowest results produced by random sampling.

Results on Tiny ImageNet. As shown in Figure 4.3(c), our PSL outperforms all other methods and achieves 17.3% accuracy with $6,000$ labelled data used. A large performance drop is observed for LL4AL, possibly because LL4AL maintains an extra loss learning module which not only trains for sampling but also updates the task learner. The limited data and huge classes might have misled the task learner through the learning module, causing the performance to be even worse than random sampling. We also notice that the random



(a) (b)

FIGURE 4.4: (a) Active learning in the sensitive protection setting on Cityscapes. The x-axis shows the counts of labelled data. (b) Active learning in the standard setting on Cityscapes. The x-axis shows the percentages of labelled data. The error bars represent the standard deviations among results from multiple different runs.

sampling baseline achieves similar results with Core-set and VAAL, possibly because Core-set relies on the quality of extracted features and VAAL relies on the training quality of VAE, which may not generalize well with only tens of annotations for each class. The results demonstrate the ability of PSL on a challenging large-scale dataset.

Results on Cityscapes. As shown in Figure 4.4(a), PSL outperforms all other methods with a clear margin. All methods start with mean IoU of 32.7% using 30 labelled data. The margin between PSL and the rest methods becomes more significant when 150 data are labelled. Eventually with 180 labelled data, PSL achieves 47.5%. Compared to the second best method LL4AL, PSL improves the results by 1.1%.

4.4.4 Active Learning in Standard Setting

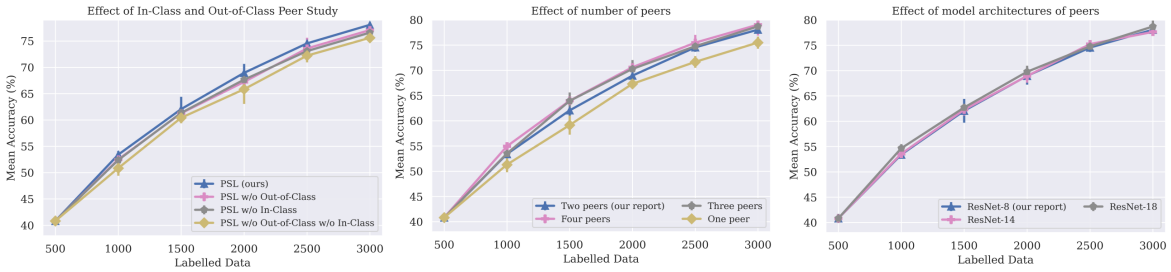
Apart from the sensitive protection setting of active learning, we also conduct experiments in the standard active learning setting where no data is protected. We use 10% of the whole dataset as the initial labelled pool and set the budget as 5% of the whole dataset. Again, we conduct the experiments on CIFAR10/100, Tiny Imagenet, and Cityscapes. We reran the models with ResNet-18 backbone for classification and DRN backbone for segmentation under the same settings for a fair comparison. Results are shown in Figure 4.3(d)-(f) and

Figure 4.4(b). In the standard active learning setting, our PSL still outperforms all the rest methods, demonstrating that our PSL is suitable for not only the sensitive protection setting, but also the standard setting of active learning. We also want to clarify that, although the standard setting for active learning does not protect specific data in the unlabelled pool as the sensitive protection setting does, our PSL still to some extent protects the data privacy by only uploading partial data to the cloud and keeping the rest to the local device.

4.4.5 Ablation Studies

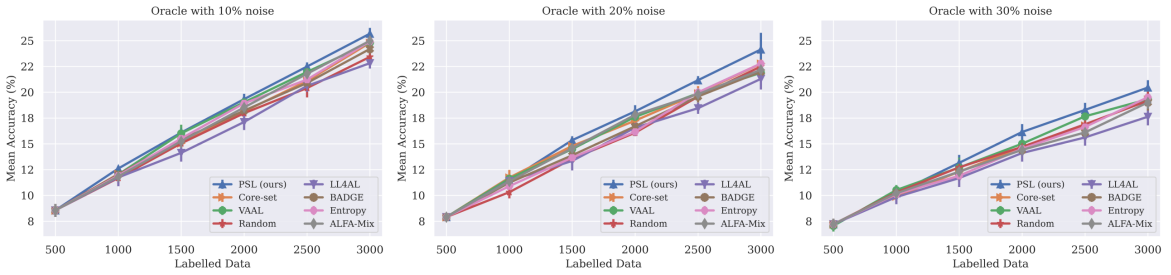
Effect of In-Class and Out-of-Class Peer Study. In Figure 4.5(a), we analyze the effect of In-Class and Out-of-Class Peer Study on CIFAR10. We show the results of PSL, PSL without Out-of-Class Peer Study, PSL without In-Class Peer Study, and PSL without both In-Class and Out-of-Class Peer Study. Specifically, we observe a clear performance drop, around 1.5%, without using Out-of-Class Peer Study or In-Class Peer Study. Without both In-Class and Out-of-Class Peer Study, the performance drop is around 2.5%. These results demonstrate that both In-Class and Out-of-Class Peer Study are crucial to performance improvement in PSL. In-Class Peer Study and Out-of-Class Peer Study depend on and supplement each other in the whole process. In-Class Peer Study provides a better base for the light-weight peers. The randomness added to peers in this process also prepares for the later sampling procedure. Out-of-Class Peer Study, on the other hand, introduces a ranking strategy to improve the sampling results, further making up for the deficiency of light-weight peers. Then Peer Study Feedback actively samples data for annotation. All together make it possible to allow responsible active learning while benefiting from cloud resources.

Effect of multiple peers. In Figure 4.5(b), we study the effect of different number of peer students. Specifically, we show the results using two, three, or four peer students. We also include the result of a single student, i.e., the Peer Study Feedback sampling criterion degrades to the entropy scores. It is observed that two peer students achieve a good trade-off between performance and the costs of computation and memory, while increasing the number of peer students brings further performance improvement. The performance improvement is larger



(a) (b) (c)

FIGURE 4.5: Ablation studies conducted in the sensitive protection setting on CIFAR10.



(a) (b) (c)

FIGURE 4.6: The results from using noisy oracles on CIFAR100. The noisy oracles with 10%, 20%, 30% noise are used respectively.

when the labelled data pool is relatively small. Extra peers benefit the model performance with the cost of memory and computation cost. In real-world applications, the number of peers to use can be left to end users to decide.

Effect of peer model architectures. Since we use light-weight peer students as active learner, naturally raises the question of using much stronger students. Aside from ResNet-8 (#parameters=0.078M), we also present the results for using ResNet-14 (#parameters=0.18M) and ResNet-18 (#parameters=11.17M) as student networks in Figure 4.5(c). We observe that ResNet-14 produces similar results compared to ResNet-8, and ResNet-18 outperforms the other two methods with a small margin. This demonstrates that using ResNet-8 as peer students is a cost-efficient choice. The improvement from using ResNet-18 for peer students is costly, with the model size 143 times of ResNet-8's size and 62 times of ResNet-14's size.

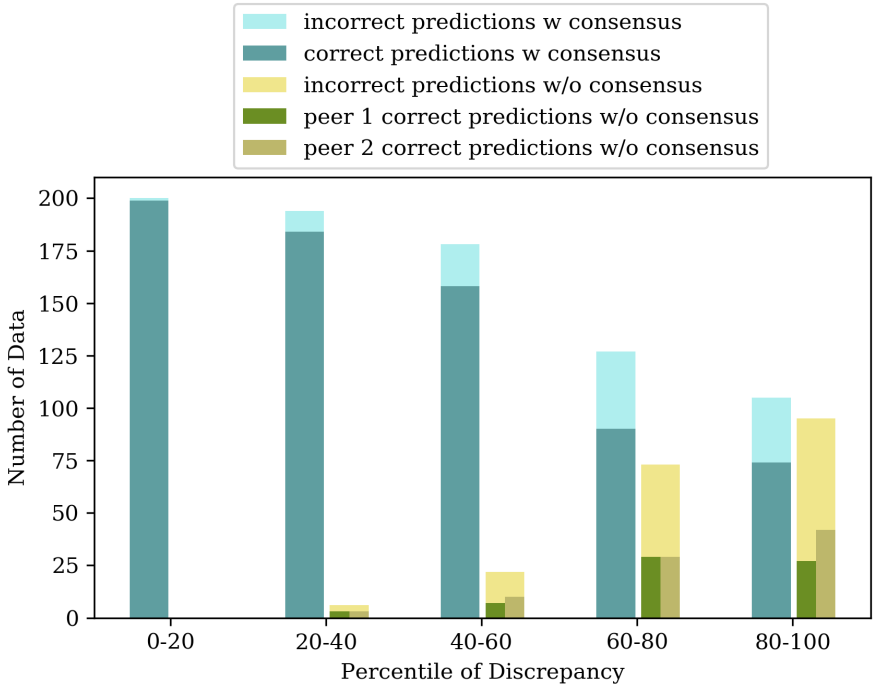


FIGURE 4.7: A further justification of Out-of-Class Peer Study on CIFAR10. Specifically, data that reach consensus between peers and data that do not reach consensus between peers are shown in blue and green respectively. Darker colours represent correct predictions and lighter colours represent incorrect predictions.

Effect of a noisy oracle. In Figure 4.6, we explore the robustness of our PSL against a noisy oracle, i.e., noisy labels are included. To better model real-world label noise in human annotations, we add 10%, 20%, and 30% label noise to both the labelled pool and the oracle annotations in CIFAR100. Since the 100 classes in CIFAR100 can be grouped into 20 superclasses, it is possible to add noise within each superclass to resemble real-world scenarios where similar classes may confuse the annotators. We follow the experiment design of [148]. We observe that the performances drop for all methods with all three degrees of label noise. And of course, with a larger degree of label noise, the performance becomes worse. Our PSL outperforms all other methods with a clear margin.

Justification of Out-of-Class Peer Study. To further justify the effectiveness of Out-of-Class Peer Study, we also perform the following experiments. As shown in Figure 4.7, we train the model with In-Class Peer Study using $4,000$ data and test it with $1,000$ data on

CIFAR10. We sort the testing data's discrepancies between peer students 1 and 2 in ascending order and divide them evenly into five groups. In each group, the predictions with consensus always have a higher accuracy than the predictions disagreed between students. This indicates that when we sample with discrepancy, most of the data selected are those cannot be correctly predicted and can improve our target network. Our Out-of-Class Peer Study aims at pushing the data with predictions reaching consensus among students to have a lower discrepancy, and thus making these data less likely to be selected during active sampling. Meanwhile, we increase the model discrepancy for the data with no consensus between students, to increase their active sampling priority.

Federated learning via responsible active learning. As we mentioned in a previous part of this chapter (subsec 4.3.5), we could fit our responsible active learning in the federated learning framework. Multiple local users are involved in this setting where each of them separately stores part of the training data, e.g., each user can be a photo album on a personal device, and data of different users can be utilized securely and efficiently to build a powerful model on the cloud. Note that this differs from the previous multiple-peer experiment in Subsection 4.4.5 which involves only one local device despite having multiple peers as the active learner. To run the experiments, we simply split CIFAR10 into random splits to distribute on 10 local clients. On each client, we protect 90% data just like in the sensitive protection setting. Each local client is composed of one active learner (peer students), and one helper model that can be uploaded for aggregation. A task model on the server is maintained to aggregate the local helper models. The peer student, the helper models, and the task model all have the same architecture, ResNet8. The peer students can access all training data and they conduct active sampling. The helper model is trained with only the data sampled from the non-protected set that are considered to be less sensitive. All helper models are uploaded and aggregated using FedAvg [115]. The aggregated model is then downloaded to initialize each helper model. We train each client for 40 epochs in each communication round and run for 50 communication rounds in total. This federated learning via responsible active learning framework allows us to strictly protect the extremely sensitive data and also avoid uploading raw training data. We run this experiment with our PSL and several state-of-the-art baselines

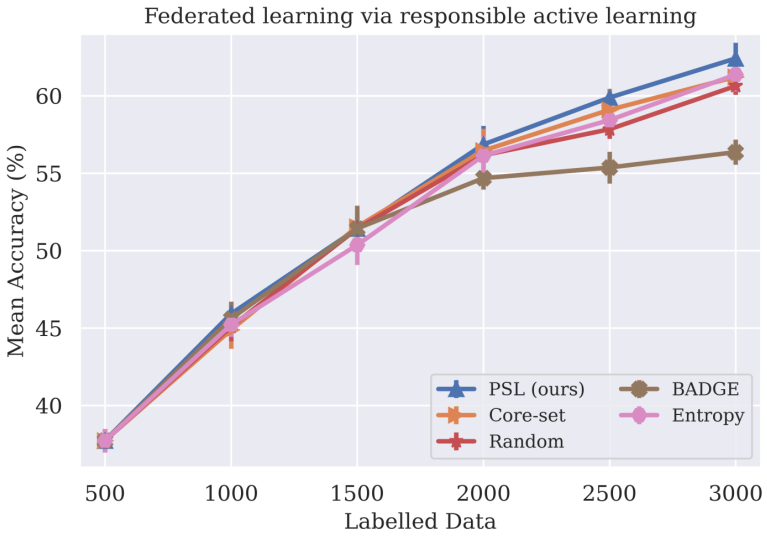


FIGURE 4.8: The results of federated learning via active learning on CIFAR10. Multiple local users/clients that locate separately are involved in this setting.

plus random sampling. We show the results in Figure 4.8. Our PSL outperforms the other baselines. BADGE performs quite poorly in this setting. Possibly because the limited data on each client makes it difficult to utilize the estimated gradients. When we compare these results to the results from sensitive protection setting shown in Figure 4.3(a), we find that the results are poorer. This is due to the decentralized training on separated data. More could be explored on the combination of federated learning and responsible active learning, such as the non-IID data distributions across clients. However, these are out of the scope of this work, we will leave it as the subject of future study.

4.5 Conclusion

In this chapter, we introduced a new Peer Study Learning framework for responsible active learning, which can simultaneously preserve data privacy and improve model stability. To protect the user data, a teacher-student architecture is introduced to isolate unlabelled data from the task learner by maintaining an active learner for active sampling. In addition, a new learning paradigm is devised to mimic the real-world in-class and out-of-class scenarios. Extensive evaluations have been conducted on both sensitive protection setting and standard

active learning benchmarks, where our proposed PSL achieves superior results over a wide range of state-of-the-art active learning methods. We also demonstrated model robustness with a noisy oracle and provided detailed component analysis to validate the insights of our model design.

CHAPTER 5

Knowledge-Aware Federated Active Learning with Non-IID Data

In this chapter, the study of insufficient training labels in deep learning is extended to the federated learning setting. We explain how the insufficiency of training labels can pose a serious problem in federated learning and why existing active learning methods cannot be directly utilized in this context.

Federated learning enables multiple decentralized clients to learn collaboratively without sharing local data. However, the expensive annotation cost on local clients remains an obstacle in utilizing local data. In the chapter, we propose a federated active learning paradigm to efficiently learn a global model with a limited annotation budget while protecting data privacy in a decentralized learning manner. The main challenge faced by federated active learning is the mismatch between the active sampling goal of the global model on the server and that of the asynchronous local clients. This becomes even more significant when data is distributed non-IID across local clients. To address the aforementioned challenge, we propose Knowledge-Aware Federated Active Learning (KAFAL), which consists of Knowledge-Specialized Active Sampling (KSAS) and Knowledge-Compensatory Federated Update (KCFU). Specifically, KSAS is a novel active sampling method tailored for the federated active learning problem, aiming to deal with the mismatch challenge by sampling actively based on the discrepancies between local and global models. KSAS intensifies specialized knowledge in local clients, ensuring the sampled data is informative for both the local clients and the global model. Meanwhile, KCFU deals with the client heterogeneity caused by limited data and non-IID data distributions by compensating for each client's ability in weak classes with the assistance

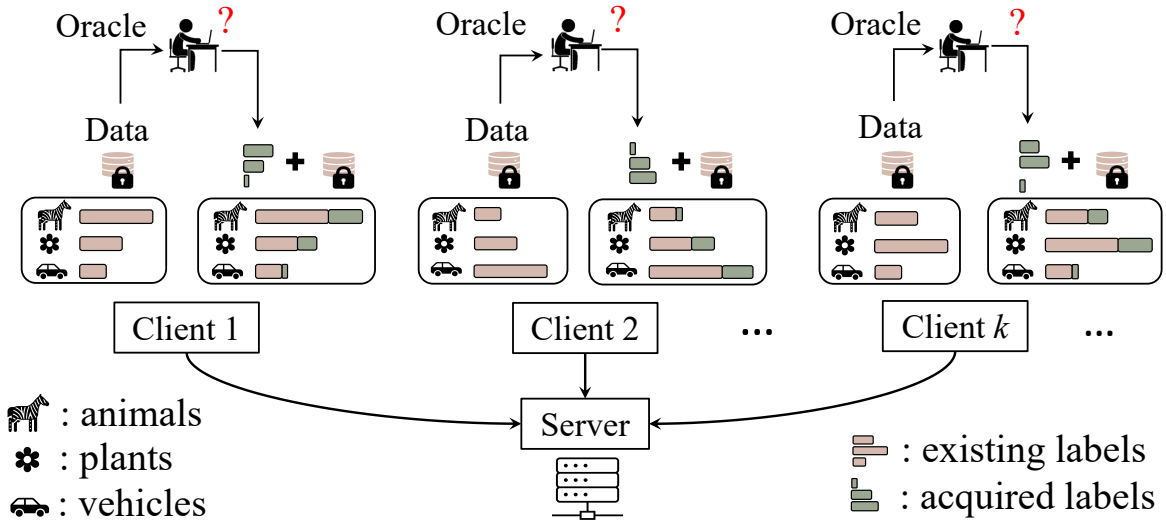


FIGURE 5.1: The primary federated active learning framework with non-IID data. Each client maintains an active learning loop to select informative data for annotation with a limited annotation budget. We show each model's existing labelled data in different classes with pink bars and the newly acquired labels with green bars. Clients specialize in different classes due to non-IID data distributions.

of the global model. Extensive experiments and analyses are conducted to show the superiority of KAFAL over recent state-of-the-art active learning methods.

5.1 Introduction

Federated learning is a decentralized paradigm that allows collaborative learning of local devices to attain a powerful global model in a central server through aggregation without accessing local data [79, 115]. Most federated learning methods consider supervised learning scenarios with fully annotated training data on each local client. However, the high annotation cost has been a challenge for real-world federated learning scenarios, e.g., large-scale medical data located in different medical institutions while medical specialists for data annotation are very limited in each institution. In this work, we consider a new federated active learning paradigm, which aims to not only protect data privacy but also make the most of the very limited annotation budget on each local client for decentralized model training. An illustration of the proposed federated active learning with non-IID data framework is shown in Fig. 5.1.

In federated active learning, we aim to attain a powerful global model on the server by sampling only local data and training the model on each local client. A straightforward solution for federated learning is to directly apply off-the-shelf active learning methods to each client. Specifically, existing methods can mainly be categorized into diversity-based [140, 3], uncertainty-based [185, 148, 192, 170, 75, 34, 5], and discrepancy-based. [141, 26]. Therefore, we actively sample based on either the statistics from each client model or the downloaded global model. However, the former approach may yield benefits primarily for local clients, while the latter might result in the loss of valuable information during aggregation, even if the selected data is advantageous for the global model on the server. Through experiments in subsequent sections, we demonstrate that active sampling with the global model on the server struggles to derive benefits due to this indirect process.

A major challenge in federated active learning is therefore, the mismatch between the active sampling goal of the clients and that of the model on server caused by asynchronous models. What makes it even more challenging is the statistical heterogeneity resulting from the non-IID data distributions on clients in a typical federated learning setting [115, 197, 63, 94]. Ideally, the models can synchronize with sufficiently many aggregations from local clients to the model on the server. However, the communication costs usually make the above-mentioned solution impractical [115]. Therefore, the model parameters of each client and the global model vary due to non-IID distributions, leading to a higher degree of mismatch between the sampling goals.

To address the aforementioned challenge, we propose a federated active learning scheme, namely **Knowledge-Aware Federated Active Learning (KAFAL)**. It comprises two key components, Knowledge-Specialized Active Sampling and Knowledge-Compensatory Federated Update. **Knowledge-Specialized Active Sampling (KSAS)** is a new active sampling strategy, where each client model learns to intensify its specialized knowledge in order to annotate universally informative data that benefit both the clients and the global model. Specifically, we compute the intensified discrepancy between the client and global model outputs based on the specialized knowledge of each client. In addition, the insufficiency of labelled training data together with the statistical heterogeneity caused by non-IID data can degrade

the federated update quality, e.g., clients may perform weakly for certain classes. Aggregating these clients, extra communications are required to achieve convergence. Therefore, we further devise a new update rule, **Knowledge-Compensatory Federated Update (KCFU)**, by compensating for weak classes (or low-frequency classes) on each client through knowledge distillation from the global model. The main contributions of this work are as follows:

- We explore a rarely studied problem, federated active learning with non-IID data, which aims at efficiently learning a global model with a limited annotation budget on each client under a heterogeneous federated learning framework. Notably, we reveal the main challenge in federated active learning is the mismatch between the active sampling goal of the clients and that of the server caused by asynchronous models.
- We introduce a federated active learning paradigm, known as KAFAL, with a novel active sampling method KSAS and a novel federated update method KCFU to handle the aforementioned challenge. KSAS is designed to sample universally informative data by computing the intensified discrepancies between the clients' and the global model's outputs based on the specialized knowledge of each client. KCFU is devised to deal with data heterogeneity by compensating for weak classes using knowledge distillation from the global model.
- We conduct extensive experiments on different benchmarks to demonstrate the superiority of the proposed method, where comprehensive ablation studies are also provided to validate the design of the proposed KAFAL.

5.2 Related Work

5.2.1 Federated Learning

Federated learning is a learning paradigm that allows decentralized training of a model on the central server with training data distributed over a number of local clients in a non-IID manner [79, 115, 65, 122, 64, 14, 100, 49]. Specifically, Konevcny et al. [79] first introduced the term and proposed a method, FedAvg, to aggregate the client models, which was

later improved by FedAvgM to accumulate model updates with momentum [65, 64]. Federated learning has also been discussed in more practical views, such as federated multi-task learning [114], federated domain adaptation [130, 183], federated continual learning [186], semi-supervised federated learning [72, 174], and unsupervised federated learning [108]. Specifically, Jeong et al. [72] considered the deficiency of data labels in federated learning and proposed a semi-supervised solution. Ahn et al. [1] and Kim et al. [76] discussed a federated active learning paradigm, while they only considered the less realistic IID data scenario. To the best of our knowledge, we are the first to explore the active data sampling problem in the non-IID federated learning framework.

5.2.2 Active Learning

Existing active learning methods can be categorized into diversity-based [140, 3, 128] , uncertainty-based [185, 148, 192, 170, 34], and discrepancy-based methods [141, 39, 27, 116, 26] , as introduced in Chapter 6.

Many recent methods have also been proposed to enable active learning in more challenging settings, e.g., low-budget active learning [111], biased-data active learning [51], semi-supervised active learning [45, 66] and cross-domain active learning [40, 110]. Our work also considers applying active learning in a more practical decentralized federated learning setting where local data privacy is protected. Chen et al. [17] proposed a novel automated learning system for distributed active learning that requires a shared labelled set. Furthermore, Goetz et al. [48] considered active learning in a federated learning framework that studies how to select clients actively. Our work, instead, considers the active sampling of data on each local client in federated learning.

5.3 Methods

In this section, we first describe the problem setting of federated active learning and then introduce two main components, i.e., KSAS and KCFU in the proposed KAFAL.

Algorithm 3 Knowledge-Aware Federated Active Learning

Data: local datasets $\{\mathcal{D}_k^L\}_{i=1}^K$ and $\{\mathcal{D}_i^U\}_{i=1}^K$
Input: T, R, sampling budgets $\{b_i\}_{i=1}^K$
Parameter: $\mathbf{\Omega}$, $\{\boldsymbol{\omega}_i\}_{i=1}^K$

for active round a=1 to A **do**
  Federated Update: KCFU
  Initialize the global model with $\mathbf{\Omega}^0$
  for communication round $t = 1$ to T **do**
    $S_t \leftarrow$ Random subset of $\lceil R \cdot K \rceil$ clients.
    for client $k \in S_t$ **do**
      Download the global model's parameters $\mathbf{\Omega}^t$
      Copy to the client $\boldsymbol{\omega}_k^t \leftarrow \mathbf{\Omega}^t$
      $\boldsymbol{\omega}_k^{t+1} \leftarrow$ LocalUpdate($\boldsymbol{\omega}_k^t; \mathcal{D}_k^L, \mathcal{D}_k^U$)
      Upload the local model parameters to the server
    end for
    for client $k' \notin S_t$ **do**
      Keep the client model unchanged $\boldsymbol{\omega}_{k'}^{t+1} \leftarrow \boldsymbol{\omega}_{k'}^t$
    end for
    Aggregate the clients with Eq. (5.7) to update $\mathbf{\Omega}^{t+1}$
    for each client $k \in S_t$ **do**
      Download $\mathbf{\Omega}_k^{t+1}$ and save as $\hat{\mathbf{\Omega}}_k$
    end for
  end for
  Active Sampling: KSAS
  for client $i = 1$ to K **do**
    for each unlabelled data $x \in \mathcal{D}_i^U$ **do**
      for class $y \in \mathbb{C}$ **do**
        Compute $P_y^i(\boldsymbol{x})$ on class y using Eq. (5.1)
        Compute $Q_y^i(\boldsymbol{x})$ on class y using Eq. (5.2)
      end for
      Compute $D^i(\boldsymbol{x})$ using Eq. (5.3)
    end for
    Send b_i unlabelled data points with the largest D to the oracle for annotation
    Remove the annotated data in $\mathcal{D}_i^U$ and add to $\mathcal{D}_i^L$
  end for
end for
Return $\{\mathcal{D}_i^L\}_{i=1}^K$ and $\{\mathcal{D}_i^U\}_{i=1}^K$

5.3.1 Problem Setting

We illustrate the overview of the federated active learning framework in Fig. 5.1 and sum up the proposed KAFAL algorithm in Alg. 3. In federated active learning, we keep K local

client models parametrized with $\{\boldsymbol{\omega}_i\}_{i=1}^K$ and one global model on central server parametrized with $\boldsymbol{\Omega}$. Each client model i is optimized using its local training dataset $\mathcal{D}_i$. Different from standard federated learning, federated active learning annotates a subset of data samples on each client with a local active learning loop. The training set $\mathcal{D}_i$ for client i is divided into a labelled set $\mathcal{D}_i^L$ and an unlabelled set $\mathcal{D}_i^U$. In each communication round, a fraction $R \in (0, 1]$ of the total K clients are first randomly selected as a subset S_t, which simulates the real-world scenarios that some local devices may be offline from time to time. After that, the selected clients first download $\boldsymbol{\Omega}$ from the server to initialize $\{\boldsymbol{\omega}_k\}_{k\in S_t}$, and then conduct local update based on $\{\mathcal{D}_k\}_{k\in S_t}$. The updated $\{\boldsymbol{\omega}_k\}_{k\in S_t}$ will be uploaded to the central server and aggregated to update $\boldsymbol{\Omega}$. The training process terminates after T communication rounds. After that, a batch of unlabelled data is sampled from each $\mathcal{D}_i^U$, sent to the local oracle for annotation, and added to the labelled data pool $\mathcal{D}_i^L$ for each client i. The sampling budget for the client i is b_i. The active sampling process is repeated for A times, where A is set according to need. The training sets $\{\mathcal{D}_i\}_{i=1}^K$ follow non-IID distributions. All client models share the same architecture with the global model to synchronize model parameters between the client and server. Just like in federated learning, transferring the local data $\{\mathcal{D}_i\}_{i=1}^K$ across clients (or server) is prohibited in federated active learning. The objective of federated active learning is to actively annotate local data with limited budgets to improve the overall model performance without violating data privacy.

5.3.2 Knowledge-Specialized Active Sampling

Given the mismatch problem in federated active learning, informative data on each client may not be that informative to the global model due to the non-IID data distributions, meaning that using only one of them for active sampling is therefore not reliable. Computing the model discrepancy between each client and the global model allows us to consider both aspects in active sampling. But alone is insufficient. Data from rare classes in each local dataset can cause large discrepancies between the client and the global model. However, they are usually uninformative to the global model and can hardly contribute to the client model's updates. Being rare locally makes their contributions limited in the gradient computation.

Furthermore, during aggregation, the global model may not find them as informative as they are to the clients. Hence, we propose to enable each client to intensify its specialized knowledge (common class knowledge) in the computation of discrepancy to sample more informative data containing specialized knowledge. We introduce the Knowledge-Specialized KL-Divergence as follows. On top of a symmetrized KL-Divergence [70, 83], our Knowledge-Specialized KL-Divergence further incorporates a Knowledge-Specialized component to accentuate each client's specialized knowledge. The Knowledge-Specialized probability of client i being predicted to class y is formulated as:

$$P_y^i(\boldsymbol{x}) = \frac{n_{i,y}^{\lambda}\text{exp}\Big(g_y(\boldsymbol{x};\boldsymbol{\omega}_i)\Big)}{\sum_{c\in\mathbb{C}} n_{i,c}^{\lambda}\text{exp}\Big(g_c(\boldsymbol{x};\boldsymbol{\omega}_i)\Big)}, \tag{5.1}$$

where $\boldsymbol{x}$ is an unlabelled data point sampled from $\mathcal{D}_i^U$, $g_y(\boldsymbol{x};\boldsymbol{\omega}_i)$ is the prediction score at the y-th class, $\mathbb{C}$ indicates the set of all classes, $n_{i,y}$ is the number of data points that belong to class y in $\mathcal{D}_i^L$, and λ is a hyperparameter which controls the knowledge-specialized level. We name $n_{i,y}^{\lambda}$ as the knowledge weight which indicates the client's knowledge in each class. Similarly, the knowledge-specialized probability of the global model predicted to be class y can be defined as:

$$Q_y^i(\boldsymbol{x}) = \frac{n_{i,y}^{\lambda}\text{exp}\Big(g_y(\boldsymbol{x};\hat{\boldsymbol{\Omega}}_i)\Big)}{\sum_{c\in\mathbb{C}} n_{i,c}^{\lambda}\text{exp}\Big(g_c(\boldsymbol{x};\hat{\boldsymbol{\Omega}}_i)\Big)}, \tag{5.2}$$

where $\hat{\boldsymbol{\Omega}}_i$ is a copy of global model parameters downloaded from the server to client i. The knowledge-specialized KL-Divergence is defined as:

$$D^i(\boldsymbol{x}) = \sum_{y\in\mathbb{C}} \Big(P_y^i(\boldsymbol{x}) \ln \frac{P_y^i(\boldsymbol{x})}{Q_y^i(\boldsymbol{x})} + Q_y^i(\boldsymbol{x}) \ln \frac{Q_y^i(\boldsymbol{x})}{P_y^i(\boldsymbol{x})}\Big), \tag{5.3}$$

where $\boldsymbol{x}$ is data from the unlabelled pool of client i. The knowledge-specialized KL-Divergence focuses on each client's specialized knowledge and selects more informative data points from its specialized classes for labelling. In the Knowledge-Specialized probabilities, $\{n_{i,c}\}_{c\in\mathbb{C}}$ serve to amplify the KL-Divergence on class c if the class is considered to contain the client's specialized knowledge.

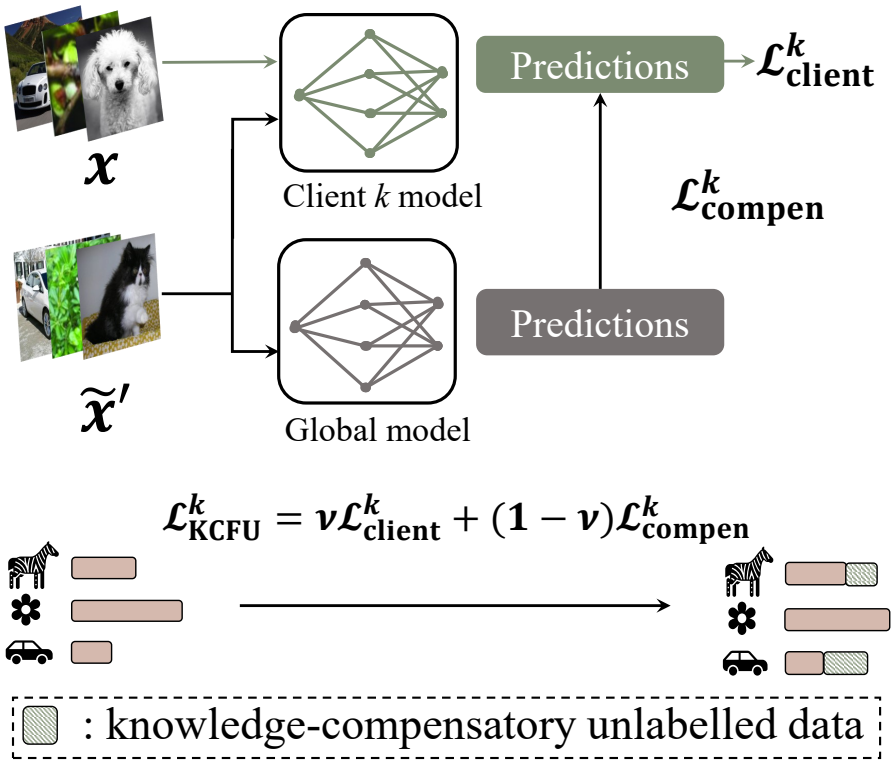


FIGURE 5.2: KCFU compensates for each client's ability on weak classes through knowledge distillation from the global model. Unlabelled data are used in the process.

5.3.3 Knowledge-Compensatory Federated Update

An overview of knowledge-compensatory federated update (KCFU) is shown in Fig. 5.2. The local data on each client follows its own realistic data distributions [65], thus leaving non-uniform class distribution on each client. Besides, our KSAS which tends to annotate data with specialized-knowledge further introduces imbalance in labelled data. Therefore, on top of the standard FedAvg [115], we introduce a balanced classifier and a knowledge-compensatory strategy.

Local Update with Balanced Loss. Our balanced classifier on each client optimizes with a balanced cross-entropy loss [135] to deal with the imbalanced local data distribution:

$$\mathcal{L}^k_{\text{client}} = -\log \frac{n_{k,y}\exp\Big(g_y(\boldsymbol{x};\boldsymbol{\omega}_k)\Big)}{\sum_{c\in\mathbb{C}} n_{k,c}\exp\Big(g_c(\boldsymbol{x};\boldsymbol{\omega}_k)\Big)}. \tag{5.4}$$

In our experiment, each labelled set $\mathcal{D}_k^L$ starts from only a small proportion of the local dataset on the client k. We demonstrate with experiments in later sections that, with the imbalanced class distributions and the small size of training data, a simple local client update using cross-entropy loss is not enough for training. The balanced loss allocates more weight to data from rare classes and less weight to data from common classes to deal with the problem. It prevents the model from becoming biased towards common classes during training.

Remark: Although our balanced loss (Eq. (5.4)) looks similar in formulation compared with the aforementioned knowledge-specialized probabilities, i.e., Eq. (5.1) and (5.2), they are designed for various purposes and function differently. Eq. (5.1) and (5.2) are designed to compute the active sampling scores, and Eq. (5.4) is a loss that updates the client models. The knowledge-specialized probabilities magnify the KL-Divergence computed from common classes for sampling, while the balanced loss magnifies the loss computed from rare classes.

Global-to-Local Knowledge Compensation. Due to the extreme limitation and non-uniformity of local data labels, the clients can perform weakly on rare classes. The weak classes of clients differ depending on the data distributions. Such statistical heterogeneity of clients can be harmful in model aggregation. To compensate for the clients' knowledge on the weak classes, we further introduce an extra loss $\mathcal{L}_{\text{compen}}$. Since the global model aggregates parameters of local clients, they usually have a more balanced performance over different classes. On classes where each client considers to be rare, the global model is likely to perform better than the client. Hence, it is reasonable to design the knowledge-compensation process which conducts knowledge distillation from the global model to the clients using unlabelled data. We later show with experiments that $\mathcal{L}_{\text{compen}}$ can save the communication cost via boosting the convergence. The loss $\mathcal{L}_{\text{compen}}$ on client k can be evaluated as follows. We first sample unlabelled data $\boldsymbol{x}'$ from $\mathcal{D}_k^U$. Then we compute the logits $\boldsymbol{z} = g(\boldsymbol{x}'; \boldsymbol{\Omega})$ and the pseudo label $y' = \arg\max_c g_c(\boldsymbol{x}'; \boldsymbol{\Omega})$ with the downloaded global model. The loss weight can be computed as $\boldsymbol{\Gamma}(\boldsymbol{x}') = \frac{\sum_{c \in \mathbb{C}} n_{k,c}}{n_{k,y'}}$, and the compensation loss is then defined as:

$$\mathcal{L}_{\text{compen}}^k = \boldsymbol{\Gamma}(\boldsymbol{x}') \cdot \text{KL}\Big(\sigma(\boldsymbol{z}) \| \sigma(g(\boldsymbol{x}'; \boldsymbol{\omega}_k))\Big), \tag{5.5}$$

Algorithm 4 LocalUpdate($\boldsymbol{\omega}_k; \mathcal{D}_k^L, \mathcal{D}_k^U$)

Data: $\mathcal{D}_k^L$, $\mathcal{D}_k^U$
Input: epochs, batches, communication round t, learning rate η
Parameter: $\boldsymbol{\omega}_k$

for $e = 1$ to epochs **do**
  for $b = 1$ to batches **do**
    Sample a batch $\{(\boldsymbol{x}, y)\} \subseteq \mathcal{D}_k^L$
    Compute $\mathcal{L}_{\text{client}}^k$ using Eq. (5.4)
    if t equals 1 **then**
      $\boldsymbol{\omega}_k \leftarrow \boldsymbol{\omega}_k - \eta \nabla \mathcal{L}_{\text{client}}^k$
    else
      Sample a batch $\{\boldsymbol{x'}\} \subseteq \mathcal{D}_k^U$
      Construct a mixed batch $\{\tilde{\boldsymbol{x}}'\}$
      Find $\boldsymbol{z}$, $\boldsymbol{\Gamma}(\tilde{\boldsymbol{x}}')$ for each $\tilde{\boldsymbol{x}}'$
      Compute $\mathcal{L}_{\text{compen}}^k$ using Eq. (5.5)
      Compute $\mathcal{L}_{\text{KCFU}}^k$ with Eq. (5.8)
      $\boldsymbol{\omega}_k \leftarrow \boldsymbol{\omega}_k - \eta \nabla \mathcal{L}_{\text{KCFU}}^k$
    end if
  end for
end for
Return $\boldsymbol{\omega}_k$

where σ stands for the softmax function and KL-divergence $\text{KL}(p, q) = p \ln \frac{p}{q}$. Note that no gradient is computed for the global model Ω, only. As the unlabelled data falls in the same distribution as the labelled data, rare classes in labelled data are usually still rare in unlabelled data. To make the most of the compensation loss, we further propose to augment the training with mixed unlabelled data $\tilde{\boldsymbol{x}}' = \boldsymbol{\beta}\boldsymbol{x}'_1 + (\mathbf{1} - \boldsymbol{\beta})\boldsymbol{x}'_2$, where $\boldsymbol{\beta}$ is a mixing weight sampled from a beta distribution. $\boldsymbol{x}'_1$ and $\boldsymbol{x}'_2$ are randomly sampled from the unlabelled batch. $\Gamma(\tilde{\boldsymbol{x}}')$ is similarly mixed as $\boldsymbol{\beta}\Gamma(\boldsymbol{x}'_1) + (\mathbf{1} - \boldsymbol{\beta})\Gamma(\boldsymbol{x}'_2)$. The compensation loss then becomes $\mathcal{L}_{\text{compen}}(\tilde{\boldsymbol{x}}'; \tilde{\boldsymbol{z}}, \boldsymbol{\omega}_k)$ with $\tilde{\boldsymbol{z}} = g(\tilde{\boldsymbol{x}}'; \boldsymbol{\Omega})$. Therefore, the complete loss $\mathcal{L}_{\text{KCFU}}^k$ to update client k is:

$$\mathcal{L}_{\text{KCFU}}^k = \nu \mathcal{L}_{\text{client}}^k + (1 - \nu)\mathcal{L}_{\text{compen}}^k, \tag{5.6}$$

where ν is a tradeoff hyperparameter. We show the detailed local update algorithm in Alg. 4.

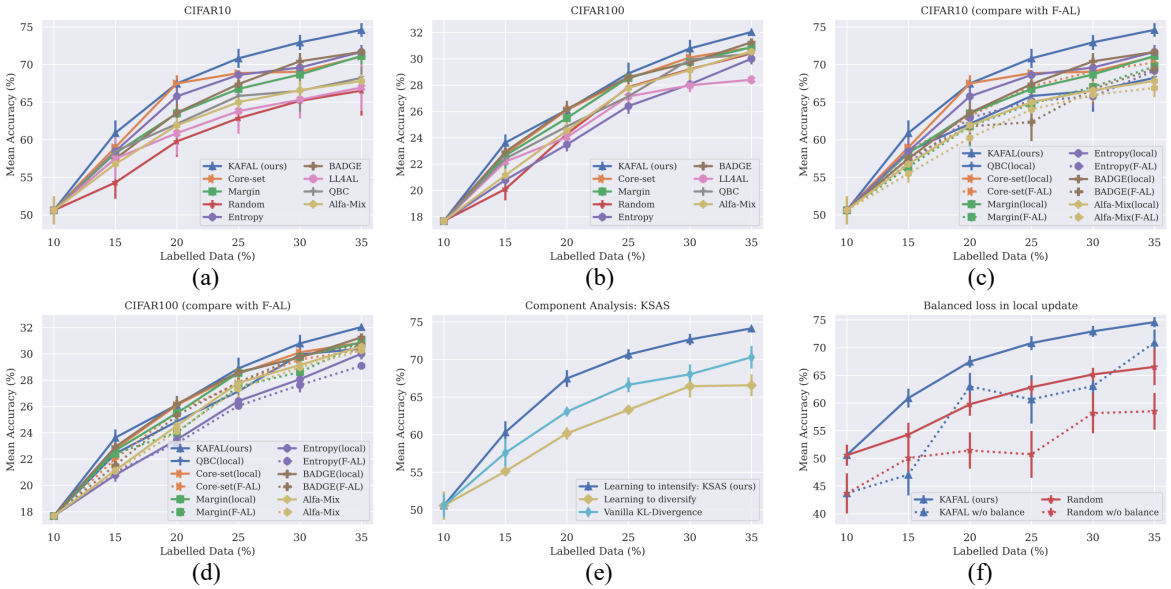


FIGURE 5.3: (a)-(b) The federated active learning results from different active learning baselines plus the results of our KAFAL on CIFAR10/100 with $\alpha = 0.1$. (c)-(d) Comparing our KAFAL with the federated active learning baseline F-AL on CIFAR10/100 with $\alpha = 0.1$. (e) Component analyses of KSAS and KCFU on CIFAR10. (f) Results with balanced loss (Eq. (5.4)) and standard cross-entropy loss on CIFAR10. For all figures, the error bars show the standard deviation of results across 5 runs.

Global Aggregation. After local updates of clients, they are uploaded to the server and aggregated as follows:

$$\boldsymbol{\Omega}^t = \sum_{k \in S_t} \frac{N_k}{\sum_{j \in S_t} N_j} \boldsymbol{\omega}_k^t, \tag{5.7}$$

where $N_k = \sum_{c \in \mathbb{C}} n_{k,c}$ indicates the number of data points in local labelled data pool $\mathcal{D}_k^L$. Lastly, we can formulate the overall objective as:

$$\underset{\{\boldsymbol{\omega}_k\}_{k \in S_t}, \boldsymbol{x} \sim \{\mathcal{D}_k^L\}_{k \in S_t}, \boldsymbol{x}' \sim \{\mathcal{D}_k^U\}_{k \in S_t}}{\arg\min} \mathcal{L}_{\text{KCFU}}, \tag{5.8}$$

where $\mathcal{L}_{\text{KCFU}} = \sum_{k \in S_t} \frac{N_k}{\sum_{j \in S_t} N_j} \mathcal{L}_{\text{KCFU}}^k$, $\{\mathcal{D}_k^L\}_{k \in S_t}$ is achieved via active learning loops.

5.4 Experiments

In this section, we mainly conduct experiments on three image classification datasets, CIFAR10/100 [81] and MNIST [89], that are popular in both active and federated learning. Additionally, we also apply our method in a more realistic scenario by conducting medical image classification with NIH Chest X-Ray dataset [172]. Specifically, CIFAR10 and CIFAR100 contain $60,000$ images from 10 and 100 classes, respectively, including $50,000$ training images and $10,000$ testing images. On CIFAR datasets, for the server and client models, we utilize ResNet-8 [58] as the model architecture. MNIST is a 10-class image dataset that contains handwritten images of 10 digits. It contains $60,000$ images for training and $10,000$ for testing. We use the MNIST 2NN proposed by McMahan et al. [115] as the clients' and the global model's architecture. We train for 10 epochs each communication round and we repeat for 10 communication rounds. We split the MNIST dataset with $\alpha = 1$.

We implement the method with Pytorch [129]. We use $K = 10$ clients in our experiments. In each communication round, $R = 80\%$ of the clients are selected at random to update locally. The hyperparameter λ is set to 1. To distribute non-IID data to different clients, we follow Hsu et al. [65] and draw $q \sim \text{Dir}(\alpha p)$ from a Dirichlet distribution. p stands for the global prior class distribution over all classes, and $\alpha > 0$ is a concentration parameter that controls the level of IID. When $\alpha \to \infty$, the data distributions are identical to the global class distribution. When $\alpha \to 0$, each client will be allocated data from only one class. In our main results, we set $\alpha = 0.1$. We also show the results from $\alpha = 0.3$ and $\alpha = 1$.

For active learning loops, we start by randomly selecting 10% data from $\mathcal{D}_i$ as the labelled pool $\mathcal{D}_i^L$ of client i. This is around 500 labelled data for each client. For each sampling cycle, the budget b_i on each client is 5% of the total local data $\mathcal{D}_i$. We sample for $A = 5$ times until the labelled data amount reaches 35% of all data for each client. We repeat each experiment 5 times with different random seeds and average the results to get a final result.

5.4.1 Comparison with Active Learning Methods

We compare our KAFAL with 8 other active learning methods and show the results in Fig. 5.3(a)(b). All methods are fit into the federated active learning framework using the same model architectures following the same training steps for fair comparison. For all baselines, we use the KCFU loss (Eq. (5.6)) for local update. FedAvg is used for aggregation for all methods. We categorize our baselines into five types. (I) We compare with uncertainty-based methods: entropy and top-2 margin scores (Margin). Entropy is calculated as $H(p) = -p \cdot \log(p)$, where p is the Softmax output. The top-2 margin score calculates the margin between the largest prediction score and the second largest prediction score over all classes for each data point. Unlabelled data with the lowest top-2 margin scores will be sampled for annotation. Here we compute the uncertainty scores on each client model after local update. (II) We compare with a special uncertainty-based method which explicitly learns the data loss with extra modules, Learning Loss for Active Learning (LL4AL) [185]. We train a loss prediction module for each client model. (III) We also compare with diversity methods: Core-set [140], BADGE [3], and ALFA-Mix [128]. We sample on each client model using diversity. (IV) Results from a previous discrepancy-based method, Query-by-Committee (QBC) [27], is also compared with. We use 3 models on each client for QBC. (V) Finally, we compare with random sampling results.

On both CIFAR10 and CIFAR100, our KAFAL achieves state-of-the-art results (Fig. 5.3(a)(b)). The margins between KAFAL and other baselines become larger with the increase of labelled data. On CIFAR10, our method eventually achieve a margin of around 3% compared to BADGE, Entropy, Margin, and Core-set. LL4AL, although quite competitive in standard active learning, does not perform well in the federated active setting. Besides, LL4AL and QBC update extra model parameters of sizes $0.015M$ and $0.156M$ for each client, when each client's model size for the rest methods is only $0.078M$. On CIFAR100, the margins are less significant compared to results from CIFAR10. Probably because the 10 times of classes in CIFAR100 makes it a much more difficult task, especially consider the limited amount of labelled data for each client. Some of the methods perform poorer than Random, possibly because Random naturally diversifies in sampling. It is worth noting that in CIFAR10 and

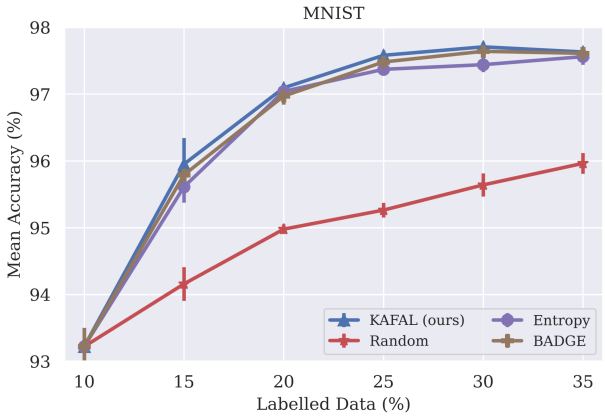


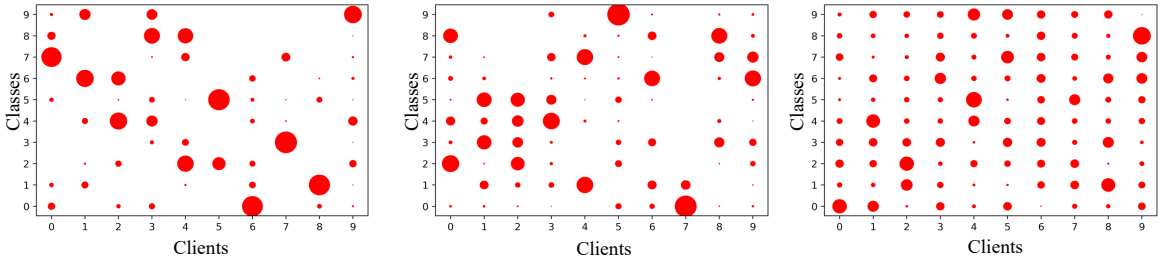
FIGURE 5.4: Results on MNIST in federated active learning with non-IID data.

CIFAR100, the full-set federated learning results are 72.93% and 37.35%. On CIFAR10, the full-set result is lower than our KAFAL result with 35% labelled data. This could be because our strategy selects only the most informative data for annotation and avoids data redundancy. On CIFAR100, the full-set result is around 5% higher than our KAFAL result with 35% labelled data, indicating that 35% data is not enough to represent a 100-class dataset. Notably, we also evaluate each client with the test set and show the analysis in later sections.

The results on MNIST are shown in Fig. 5.4. This is a fairly simple dataset, so all the results are quite high. But random is still far behind compared to the other methods. On this dataset, our KAFAL still outperforms the other methods, but with a quite small margin.

5.4.2 Comparison with Sampling by Global Model

As we mentioned in previous sections, for some sampling methods, it is possible to compute the sampling criteria either on the local client after local updates or on the downloaded aggregated global model. Using the global model for sampling is also the main idea of F-AL [1], a federated active learning method for IID data. Among the baselines we compared with in Sec. 5.4.1, we found Core-Set, Margin, Entropy, BADGE, and Alfa-Mix to be qualified to compute sampling scores on either the clients or the downloaded global model. We show the experiment results on CIFAR10 and CIFAR100 in Fig. 5.3(c)(d). The solid lines show the results from using locally updated client models to compute sampling criteria. These are



(a) $\alpha = 0.1$ (b) $\alpha = 0.3$ (c) $\alpha = 1$

FIGURE 5.5: Illustration of CIFAR10 non-IID data distributions over clients with $\alpha = 0.1$, $\alpha = 0.3$, and $\alpha = 1$. The x-axes represent the client names. The y-axes represent the class labels. The dot sizes represent the number of data.

also the results presented in Fig. 5.3(a). The dashed lines represent the results from using the downloaded global model for computing sampling scores. There is a clear drop in performance moving from client model statistics to global model statistics. Additionally, we compare our method with QBC. Our method combines local and global with discrepancy-based sampling and QBC is a local disagreement-sampling method. This experiment demonstrates the challenge in federated active learning where the sampling aims of the clients mismatch with that of the global model. It also shows that even if we sample informative data points directly using the downloaded global model, the information cannot be fully utilized to benefit the global model through aggregation. F-AL, which is initially proposed for IID federated active learning, does not suit the task of federated active learning with non-IID data.

5.4.3 Result Numbers

We present the exact result numbers of our main results on CIFAR10 (Tab. 5.1) and CIFAR100 (Tab. 5.2). To present a complete picture of our KAFAL performance, we evaluate each client using the test set and show the results in Fig. 5.6. The non-IID data distributions used in our main results on CIFAR10 are show in Fig. 5.5(a).

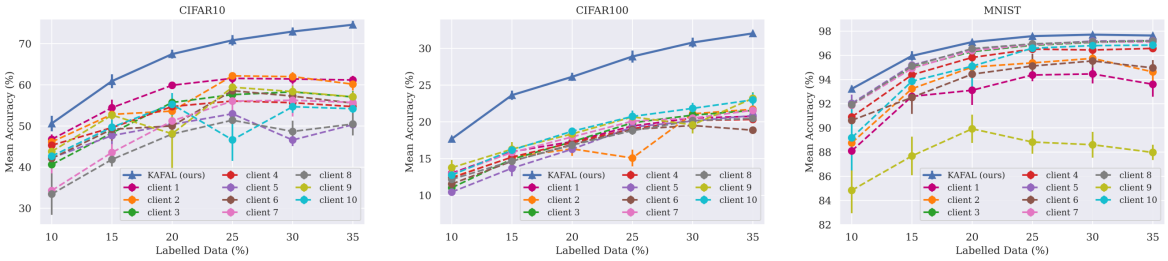


FIGURE 5.6: Evaluation on each client on CIFAR 10/100 and MNIST using KAFAL.

TABLE 5.1: Detailed results on CIFAR10.

Method	10%	15%	20%	25%	30%	35%
Random	50.60	54.29	59.76	62.85	65.16	66.52
Core-Set		58.98	67.48	68.85	69.04	71.05
Entropy		58.45	65.76	68.61	69.59	71.68
Margin		58.19	63.50	66.75	68.66	71.13
LL4AL		57.48	60.87	63.79	65.31	66.94
QBC		58.45	62.10	65.81	66.49	68.22
BADGE		57.46	63.57	67.39	70.42	71.67
Alfa-Mix		56.75	61.90	64.98	66.57	67.81
KAFAL (ours)		60.88	67.47	70.82	72.94	74.60

TABLE 5.2: Detailed results on CIFAR100.

Method	10%	15%	20%	25%	30%	35%
Random	17.67	20.10	24.28	27.85	29.21	30.41
Core-Set		22.78	26.10	28.49	30.11	30.86
Entropy		20.79	23.48	26.41	28.07	30.01
Margin		22.65	25.50	28.56	29.77	30.88
LL4AL		22.18	24.05	27.14	27.99	28.41
QBC		22.41	24.86	27.15	29.95	30.39
BADGE		22.93	26.19	28.61	29.72	31.26
Alfa-Mix		21.14	24.54	27.79	29.15	30.56
KAFAL (ours)		23.63	26.13	28.89	30.79	32.04

Pleural thickening Atelectasis Emphysema Nodule Infiltration

FIGURE 5.7: Selected images in NIH Chest X-Ray dataset.

5.4.4 Medical Image Classification

We further conduct experiments in a more realistic scenario of X-ray image classification using NIH Chest X-Ray dataset [172]. Some examples are shown in Fig. 5.7. The task is to categorize thorax diseases using chest X-ray images. The dataset consists of more than 112k images of size 1024×1024. We follow the official training and testing splits. And we exclude images tagged with 'no findings'. The rest data have 14 for different thorax diseases as labels. The training split includes 36024 images and the testing split includes 15735 images. We use ResNet-50 [57] as the backbone of the clients and the global model. We still use $\alpha = 0.1$ as the non-IID coefficient to distribute the client data. 5 clients are used, and 80% are selected for the update at each communication round. We start with 10% labels and use 5% of the whole dataset as the budget. We train for 2 epochs in each communication round with learning rate $\eta = 0.0005$ and run 5 communication rounds before sampling. The mean AUC score is used to evaluate each method's performance. The results are presented in Tab. 5.3. We compare with four baseline methods (Random, Core-Set, Entropy, and Margin) that the dataset can easily fit in considering the image size and model size. Our KAFAL still achieves state-of-the-art results on this dataset.

TABLE 5.3: mAUC scores on NIH Chest X-Ray dataset.

Method	10%	15%	20%
Random	56.12	60.62	62.77
Core-Set		62.55	63.24
Entropy		63.13	63.80
Margin		60.19	62.81
KAFAL (ours)		63.61	64.48

5.4.5 Ablation Studies

5.4.5.1 Component study

To explore the importance of our model components in KAFAL, we separately run experiments to analyze KASA and KCFU. To analyze KSAS, we first replace our knowledge-Specialized KL-Divergence with a vanilla KL-Divergence and observe a 3% to 5% performance drop

through the whole sampling process (Fig. 5.3(c)). We also attempt to sample with a reversed KSAS (learning to diversify), where we replace each $n_{i,y}$ and $n_{i,c}$ in Eq. (5.1) and (5.2) with $\frac{1}{n_{i,y}}$ and $\frac{1}{n_{i,c}}$. This prevents the client models from intensifying their specialized knowledge. Instead, it drives the clients to focus on sampling data from rare classes. The results show a significant drop compared to the other two. This further validates our design where each client should intensify its knowledge during active sampling. Data from rare classes can be quite useless in improving the global model on the server.

To analyze the efficiency of KCFU, we count the number of communication rounds of different federated update ways to achieve the same accuracy. The benchmark accuracy is set as the accuracy of running KCFU for 15 rounds. We experiment with the baseline method by replacing KCFU with a vanilla federated update which removes $\mathcal{L}_{\text{compen}}$ and updates with $\mathcal{L}_{\text{client}}$ (eq. (5.4)) only. We also compare the results from mixing and not mixing unlabelled data in KCFU. As shown in Tab. 5.4, KCFU can converge faster than vanilla update no matter mixed data are used, demonstrating the effect of our knowledge-compensatory design which borrows common knowledge from the global model. Mixing data in KCFU further boosts the efficiency. We further experiment by fixing $\mathbf{\Gamma}(\tilde{\boldsymbol{x}}') = \frac{1}{C}$, where C is the number of classes, for $\mathcal{L}_{\text{compen}}$ (Eq. (5.5)). This means we distil knowledge from the downloaded global model without differentiation on all unlabelled data. Unsurprisingly, the performance is very poor. The accuracy reaches only 40.4% with 10% data and the setting aligned with Fig. 5.3(a).

TABLE 5.4: Number of rounds of different federated update ways to achieve the same accuracy as running KCFU runs for 15 rounds.

Method	CIFAR10	CIFAR100
Baseline	35	39
KCFU w/o mix	25	27
KCFU	**15**	**15**

5.4.5.2 Local update without the balanced loss

We use a balanced loss (Eq. (5.4)) for local update of clients. This type of loss is usually the cherry on the top for standard federated learning. This is however not the case in our federated active learning problem. In Fig. 5.3(f), we show the results using balanced loss (Eq. (5.4))

and a simple cross-entropy loss (simply replacing $n_{i,y}$ and $n_{i,c}$ in Eq. (5.4) with 1). We ran the experiment with two methods, our KAFAL and random sampling. From the experiment results, we can see that removing the balanced loss in local update disturbs or almost ruins the learning, a drop of 5% to 10% in performance occurs. Our KAFAL still outperforms random sampling, but the results are highly unstable. This is somewhat foreseeable since each client model starts with a very small amount of data in federated active learning. Despite our learning to intensify on specialized knowledge during sampling, it is still crucial to handle the imbalance of data during local client update using the balanced loss.

5.4.5.3 Knowledge specialization alternatives

It is an interesting question whether other reweighting techniques can also help achieve knowledge specialization in federated active learning. Here we compare our method with two knowledge specialization alternatives, probability-level specialization and KL-Divergence-level specialization. The experimental results (Fig. 5.8) show that KAFAL outperforms both of the alternative methods. While probability-level specialization yields an acceptable outcome, KL-Divergence-level specialization fails to produce a reasonable result. One possible reason for this difference is that the probability-level specialization method, like our KAFAL, uses a moderate level of reweighting to adjust the results. In contrast, the KL-Divergence-level specialization method directly reweights the summation in the KL-Divergence calculation, potentially resulting in a level of reweighting that is too strong.

Given that knowledge specialization of KL-Divergence is achieved via score-level reweighting (as detailed in Eq. (5.1), Eq. (5.2), and Eq. (5.3)) in our KAFAL, an interesting question arises: Can other reweighting techniques also enable knowledge specialization in federated active learning? To answer this question, we compare our method with two knowledge specialization alternatives, namely probability-level specialization and KL-Divergence-level specialization.

To conduct probability-level specialization, we can rewrite Eq. (5.1) as follows:

$$P_y^i(\boldsymbol{x}) = \frac{\exp\Big(\nu_{i,y}^{\lambda} \cdot g_y(\boldsymbol{x}; \boldsymbol{\omega}_i)\Big)}{\sum_{c \in \mathbb{C}} \exp\Big(\nu_{i,c}^{\lambda} \cdot g_c(\boldsymbol{x}; \boldsymbol{\omega}_i)\Big)}, \tag{5.9}$$

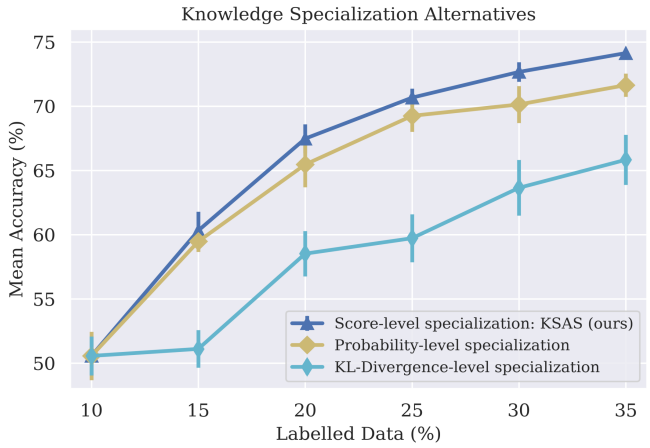


FIGURE 5.8: Results on CIFAR using different knowledge specialization techniques.

where $\nu_{i,y} = \frac{n_{i,y}}{\sum_{c \in \mathbb{C}} n_{i,c}}$ is the normalized knowledge weight. Note that we did not normalize the knowledge weight in our score-level knowledge specialization (KAFAL) because it can be easily proved that the results are equivalent with or without normalization. And similarly, Eq. (5.2) is replaced with:

$$Q_y^i(\boldsymbol{x}) = \frac{\exp\Big(\nu_{i,y}^{\lambda} \cdot g_y(\boldsymbol{x}; \boldsymbol{\Omega})\Big)}{\sum_{c \in \mathbb{C}} \exp\Big(\nu_{i,c}^{\lambda} \cdot g_c(\boldsymbol{x}; \boldsymbol{\Omega})\Big)}. \tag{5.10}$$

This knowledge specialization alternative still involves the computation of the KL-Divergence as described in Eq. (3). This knowledge specialization alternative reweights the logits during the calculation of the predicted probability, hence the name.

To conduct KL-Divergence-level specialization, we replace Eq. 5.3 with:

$$D^i(\boldsymbol{x}) = \sum_{y \in \mathbb{C}} \left[\nu_{i,y}^{\lambda} \cdot \left(P_y^i(\boldsymbol{x}) \ln \frac{P_y^i(\boldsymbol{x})}{Q_y^i(\boldsymbol{x})} + Q_y^i(\boldsymbol{x}) \ln \frac{Q_y^i(\boldsymbol{x})}{P_y^i(\boldsymbol{x})} \right) \right]. \tag{5.11}$$

This knowledge specialization alternative reweights the sum while calculating KL-Divergence.

In Fig. 5.8, we present the results of the two alternatives as well as our KAFAL. The experimental results show that KAFAL outperforms both of the alternative methods. While probability-level specialization yields an acceptable outcome, KL-Divergence-level specialization fails to produce a reasonable result. One possible reason for this difference is that

the probability-level specialization method, like our KAFAL, uses a moderate level of re-weighting to adjust the results. In contrast, the KL-Divergence-level specialization method directly reweights the summation in the KL-Divergence calculation, potentially resulting in a stronger level of reweighting. Our score-level specialization approach may outperform probability-level specialization because reweighting the raw logits may not have a natural interpretation, whereas reweighting normalized results as in our KAFAL can be interpreted as adjusting the likelihood of the results.

5.4.5.4 Different Non-IID Levels

We further explore federated active learning with the non-IID coefficient $\alpha = 0.3$ and $\alpha = 1$ on CIFAR10. The data distributions are shown in Fig. 5.5(b) and (c) respectively. We show the experiment results in Fig. 5.9. A larger α value provides less non-IID distributions for clients, i.e., the distributions across different clients are more similar. Unsurprisingly, compared to our CIFAR10 with $\alpha = 0.1$ results, the results are overall better for $\alpha = 0.3$ and $\alpha = 1$. Our KAFAL is still state-of-the-art, but the margins between the results of KAFAL and the rest methods are relatively smaller. This experiment demonstrates that our KAFAL is more competitive with higher levels of non-IID. It validates that intensifying knowledge-specialized data in KAFAL can handle the non-IID distributed data in federated active learning. The margins between Random and other methods become larger with larger α values, possibly because the mismatch problem in federated active learning becomes less significant with a lower level of non-IID in data. And the rest methods can benefit from the actively sampled data.

5.4.5.5 Different Values of λ For Knowledge-Specialized Intensification

The coefficient λ in eq. (5.1)(5.2) controls the knowledge-specialized level in KSAS. With larger values of λ, the clients intensify more on their specialized knowledge in active sampling. As we stated, we simply use $\lambda = 1$ in our main experiments. Here we explore more values of λ on CIFAR10 and show the results in Fig. 5.10. For λ of values 1, 2, and 3, the difference is not significant. However, for more extreme λ values 0.1 and 10, the results are clearly poorer.

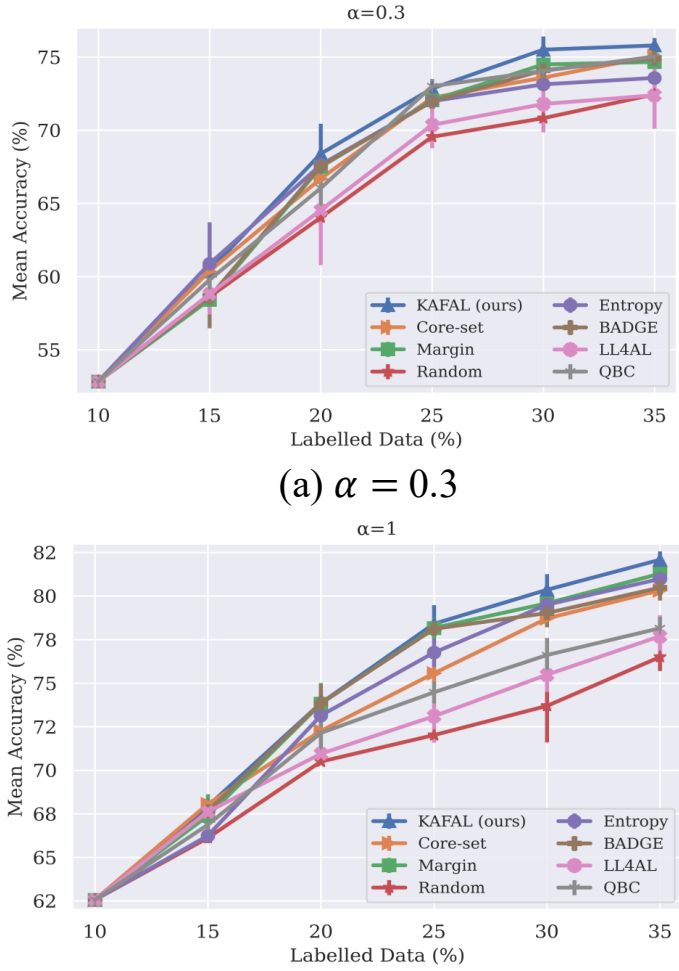


FIGURE 5.9: Results from using $\alpha = 0.3$ and $\alpha = 1$ for the non-IID coefficient on CIFAR10.

Specifically, $\lambda = 0.1$ produces the worst results of the five. When the λ value approaches zero, the active sampling purely depends on the disagreement between the clients and the global model. The results gradually approach the results from using vanilla KL-Divergence. When the λ value goes to infinity, the active sampling process almost ignores the less frequent classes and tries to compute the disagreement solely based on the most common class (or classes) of each client. Therefore, when applying KAFAL, the λ value should be neither too small nor too large.

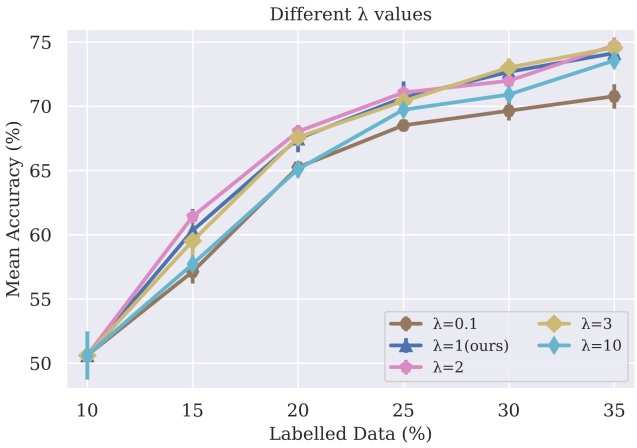


FIGURE 5.10: Results on CIFAR10 from using five different values of λ for the intensification of specialized knowledge in federated active learning with non-IID data.

5.4.5.6 Learning With More Decentralized Clients

In the main experiments, we explored federated active learning with $N = 10$ clients. To better analyze the problem, we run experiments on CIFAR10 with $N = 20$ and $N = 100$ while keeping the rest setup the same. The labelled data amount still starts with 10% of each local training set, meaning that with $N = 20$ the data available for each client is half of that in the previous experiments, and with $N = 100$ the data available for each client is only $\frac{1}{10}$ of that in the previous experiments. The results are shown in Fig. 5.11. Compared with the previous results from using $N = 10$ clients, results for all methods reduce due to the smaller local datasets for both $N = 20$ and $N = 100$. Our KAFAL still outperforms the rest methods by a clear margin. This shows the superiority of our method when more decentralized clients are involved in federated active learning. The result lines are more jiggly compared with previous results, the possible reason is that the fewer labelled data and the $T = 50$ communication rounds may not be enough for the convergence to be achieved. With $N = 100$ clients, the margin is less significant compared to using $N = 20$, this is possibly due to the extremely small local dataset size. Each local dataset starts with on average 500 images and adds about 250 images at each active round for $N = 100$. This also explains why Entropy and BADGE generate similar results compared with Random. The limited training data lead to poor classification ability and deteriorates the credibility of the model statistics.

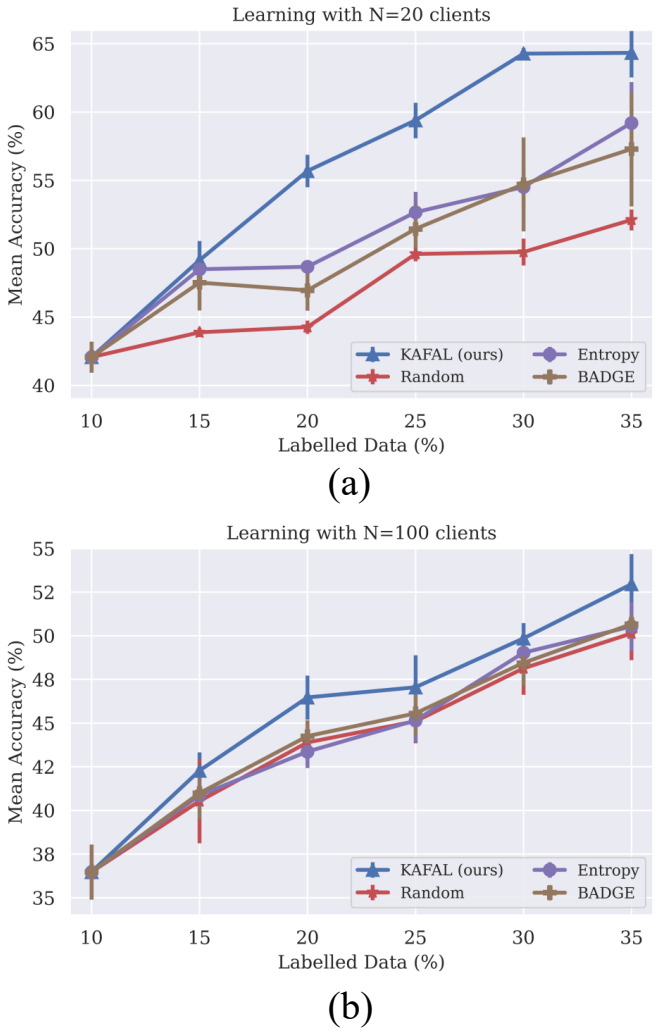


FIGURE 5.11: Results on CIFAR10 from using (a) $N = 20$ and (b) $N = 100$ clients in federated active learning with non-IID data.

5.4.5.7 A Smaller Ratio of Clients to Update per Round

We used $R = 80\%$ in previous experiments. To test how our KAFAL performs with a smaller ratio of clients updated in each communication round, we use $R = 40\%$ instead and present the results on CIFAR10 in Fig. 5.12. All the rest setup is kept the same. $R = 40\%$ means that only 40% of the clients are updated in each communication round. Surprisingly, our KAFAL performs even better using $R = 40\%$ compared with using $R = 80\%$, while results from the rest methods all drop. This is possible because our KAFAL compensates for the knowledge of clients with the global model using KCFU along with actively sampling data by intensifying

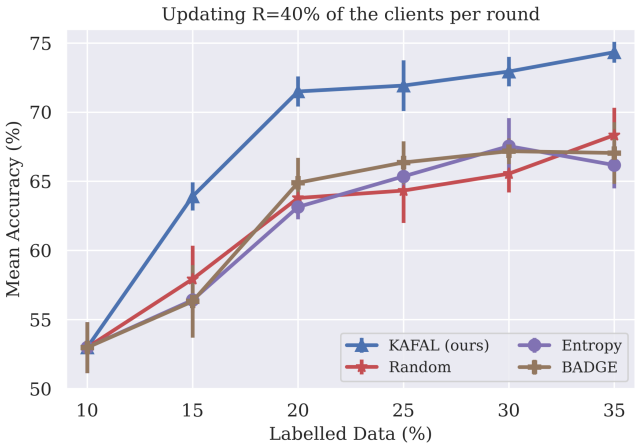


FIGURE 5.12: Results on CIFAR10 from updating 40% of clients per communication round in federated active learning with non-IID data.

specialized knowledge using KSAS. The two together enable a faster convergence in global aggregations. Using $R = 40\%$ means each client is trained less compared to using $R = 80\%$ when the communication rounds T is fixed. The rest methods which still actively sample harder data that are likely from less frequent classes cannot utilize these data in training with the smaller R value. Although KCFU is also used for other methods for a fair comparison, it cannot be fully utilized without the knowledge-specialized intensification of KSAS.

5.4.5.8 Visualizing the Mismatch Problem

Previously, we mentioned that the main challenge of federated active learning is the mismatch between the active sampling goal of the global model on the server and that of the asynchronous local clients. To demonstrate this problem with an experiment, we actively sample with the global model and the clients respectively and show the class distributions of the sampled data in Fig. 5.13. We use the same sampling method Core-set for both the clients and the global model for a fair comparison. With the bounding boxes, we show the differences between the sampling results. The original data distributions on clients with $\alpha = 0.1$ are shown in Fig. 5.5(a). Also, note that this figure only shows the class distributions. If we further consider specific data points within each class, the difference in sampled results will be more significant.

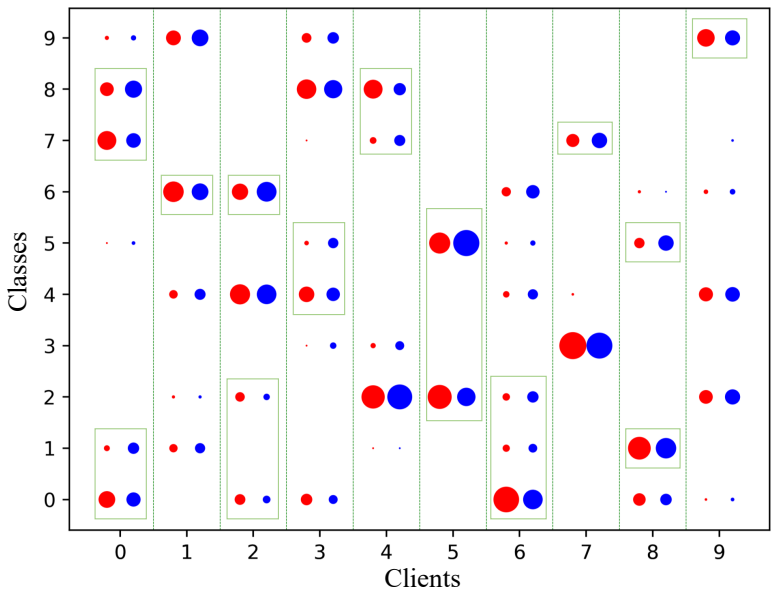


FIGURE 5.13: An example of the sampling goal mismatch between the global model and the clients. The green bounding boxes highlight class distributions that are clearly different for active sampling results using the global model (red) and active sampling results using the client model (blue).

5.4.6 Demonstration of Knowledge-Specialized KL-Divergence in a Toy Example with Details

To better visualize how Knowledge-Specialized KL-Divergence intensifies specialized knowledge compared to KL-Divergence, we use continuous distributions to simulate model predictions and compute the divergences (Fig. 5.14). Note that the knowledge weight curves serve as a continuous version of our knowledge weights. In the figure, we present the distribution curves on the left and the corresponding KL-Divergence and Knowledge-Specialized KL-Divergence curves on the right, which have been calculated accordingly. The KL-Divergence curve is formulated as:

$$p(x) \ln \frac{p(x)}{q(x)} + q(x) \ln \frac{q(x)}{p(x)}, \tag{5.12}$$

where $p(x)$ and $q(x)$ are the two distribution functions (presented on the left of Fig. 5.14). The KL-Divergence value is obtained by integrating this function with respect to x. The Knowledge-Specialized KL-Divergence curve is formulated as:

$$p_w(x) \ln \frac{p_w(x)}{q_w(x)} + q_w(x) \ln \frac{q_w(x)}{p_w(x)}, \tag{5.13}$$

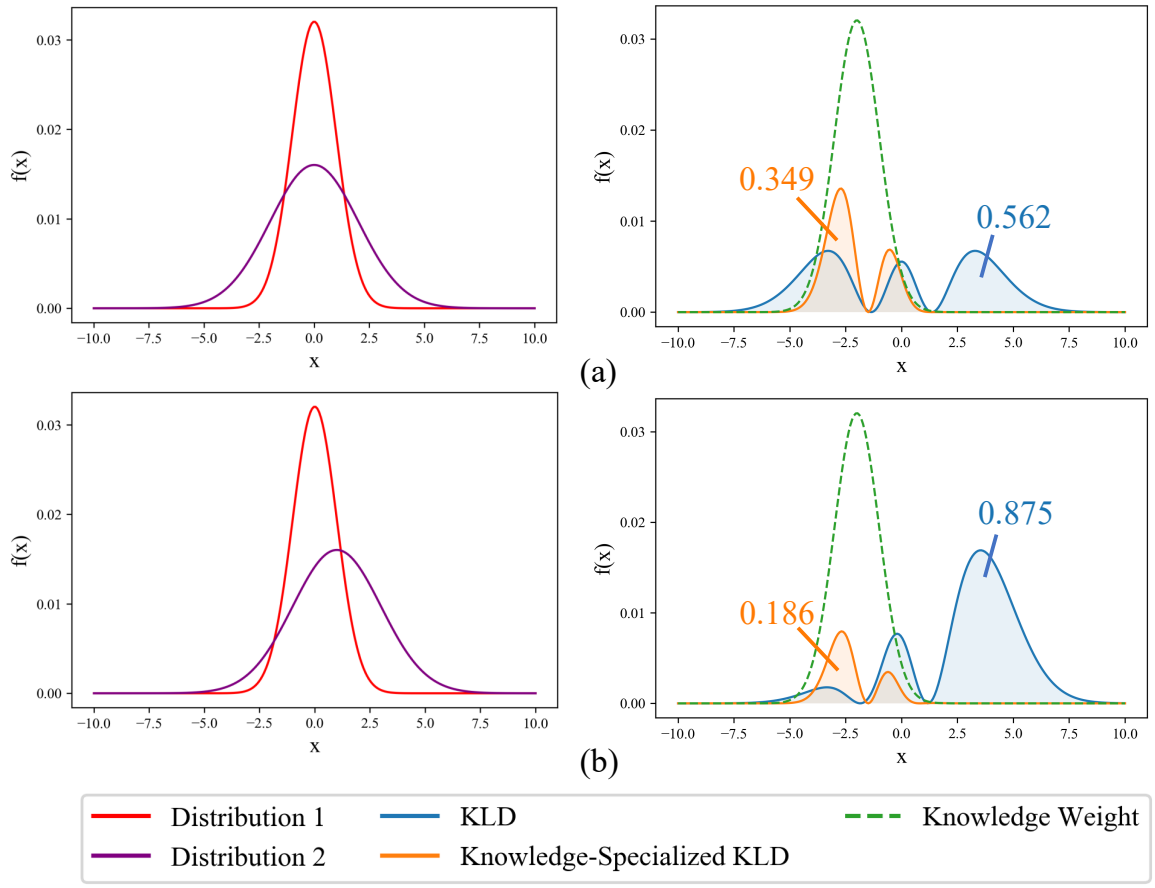


FIGURE 5.14: Illustration of how Knowledge-Specialized KL-Divergence intensifies specialized knowledge compared to standard KL-Divergence. On the left, we show two distribution curves. On the right, the blue and orange lines integrate to be KL-Divergence and the knowledge-specialized KL-Divergence computed from the left distributions. The blue and orange numbers show the integrated areas of the blue and orange curves in each image, respectively.

where $p_w(x) = \frac{w(x) \cdot p(x)}{Z_p}$ and $q_w(x) = \frac{w(x) \cdot q(x)}{Z_q}$. The normalization constants $Z_p = \int p(x) \cdot w(x) \mathrm{d}x$ and $Z_q = \int q(x) \cdot w(x) \mathrm{d}x$. The weight curve $w(x)$ is shown with green dashed lines in the figure. The Knowledge-Specialized KL-Divergence value is obtained by integrating this function with respect to x. The right-hand side of Fig. 5.14 can be viewed as global-local discrepancies from two different inputs on the same client model since the KL-Divergence values are different and the knowledge weights are the same. On the left-hand side, distributions 1 and 2 simulate the outputs of the client model and the global model. Notably, while (a) has a smaller KL-Divergence, its Knowledge-Specialized KL-Divergence is larger, suggesting

it is less likely to be sampled than (b) if KL-Divergence is the sampling criterion. However, using our proposed Knowledge-Specialized KL-Divergence, (a) is more likely to be sampled than (b). This difference in sampling results is due to the knowledge weight, which intensifies the client's specialized knowledge while dampening the contribution of unfamiliar knowledge. Importantly, in (a), more of the model difference arises from specialized knowledge (as indicated by the peak area of the knowledge weight) compared to (b).

5.5 Limitations and Future Work

Our federated active learning paradigm KAFAL includes KSAS, a novel active sampling method to sample informative data using intensified discrepancies between the server and clients based on the specialized knowledge of each client, and KCFU, a federated update method to deal with data heterogeneity by compensating weak classes with the help from the global model. Although the experimental results demonstrate that KAFAL can perform well on the federated active learning task, we also want to highlight the potential drawbacks of this method. In KSAS, the specialized knowledge is extracted based on the class distributions of labelled local data. We may explore other ways to find a more comprehensive solution to represent the specialized knowledge, either, possibly not only considering the class distributions but also taking the training dynamics into account. In KCFU, the compensation is achieved through sampling the unlabelled data and then weighting them using the class distributions. Unfortunately, the data from weak classes may not be enough even though we include the unlabelled data. We may utilize the data generation techniques to generate more weak-class data for better knowledge compensation in the future. In addition to the potential drawbacks mentioned, another area for future work is to extend KAFAL to handle the case of long-tailed distribution in the federated active learning setting. In a long-tailed scenario, the local data can distribute globally long-tailed with some classes being rare for all clients. To consider active learning in such a scenario, additional resampling techniques and an improved version of knowledge-specialized KL-Divergence that takes the long-tailed distribution into account need to be included.

5.6 Conclusion

We have introduced a federated active learning paradigm which allows actively selecting the unlabelled data to efficiently learn a global model given a limited annotation budget in a decentralized learning process. We revealed that the main challenge in federated active learning is the mismatch between the active sampling goals of the global model on the server and each local client due to model differences caused by non-IID data distributions. This work devised a Knowledge-Aware Federated Active Learning (KAFAL) method for federated active learning with non-IID data. KAFAL computes the discrepancies between client-server models with an intensification on each client's specialized knowledge. It is worth noting that the intensifying process is particularly important to achieve a powerful global model in the non-IID federated learning framework. Moreover, KAFAL also compensates for each client's ability in rare classes to handle data heterogeneity caused by non-IID data during federated updates. Extensive experiments and analyses have validated the superiority of KAFAL over the state-of-the-art active learning methods under the federated active learning framework.

CHAPTER 6

Active Human-Object Interaction Detection

This chapter will study another special case in insufficiency of training labels where the labels are in complex forms. In the previous chapters, we only studied problems where only simple-labelled data are involved, e.g., classification and segmentation. In this chapter, we propose an active learning method for human-object interaction (HOI) detection which involves identifying human-object pairs and discerning interaction types within images featuring multiple interactions and objects across diverse scenes. Annotating all combinations incurs substantial costs. While active learning presents a promising solution, applying it for HOI detection is non-trivial due to challenges such as assessing image informativeness and handling spurious features. To tackle these issues, we propose ActiveHOI, a novel approach that integrates informativeness measures from various image components (such as object categories, bounding boxes, and interaction types) while also addressing spurious background information during the active HOI detection process. Specifically, we develop a comprehensive metric to quantify the similarity between the original image and its foreground-masked counterpart, effectively guiding the active sampling process. Through this method, we aim to enhance the efficiency and effectiveness of HOI detection while reducing annotation costs. We evaluate ActiveHOI on two popular HOI detection benchmarks, HICO-DET and V-COCO, and the experimental results demonstrate its effectiveness compared to recent state-of-the-art active learning methods.

6.1 Introduction

Human-object interaction (HOI) detection is crucial in visual understanding, encompassing identifying specific interaction types between all possible human-object pairs in images. Thought recent advancements in HOI detection have been notable [13, 152, 98, 132, 91, 177], the inherent complexity of such data, rich interactions and diverse scenes, poses significant challenges, particularly regarding annotation costs. Specifically, the HOI annotation process typically involves manually annotating images with bounding boxes of humans and objects, as well as labels of detected objects and interaction labels of human-object pairs. Furthermore, due to the combinatorial nature, annotating all possible human-object pairs within each image to construct a large-scale training dataset containing images with multiple humans and objects in complex scenes becomes labor-intensive. Therefore, identifying cost-effective annotation strategies is crucial for advancing HOI detection systems.

Active learning efficiently manages data annotation costs by identifying high-priority informative instances from the pool of unlabeled data [42, 185, 148, 192, 5]. Consequently, models trained with limited annotation budgets are expected to perform comparably to those trained with annotation on the entire dataset. Existing active learning methods fall into three categories: 1) model uncertainty-based [42, 185, 148, 192, 5, 128, 170, 75, 34]; 2) model discrepancy-based [141, 39, 22, 116, 27, 26]; and 3) data diversity-based [140, 3]. However, applying active learning techniques to HOI detection is non-trivial due to the following challenges. Specifically, active image classification, a common application of active learning, typically deals with images containing a single prominent item positioned centrally, reflecting the nature of classification tasks. In contrast, HOI detection involves analyzing images with complex scenes, where multiple objects or humans are dispersed across varying sizes and locations within the image. A similar task to refer to would be the active object detection problem since it also involves image data with complex scenes. While several existing works [20, 190, 117] in the literature address the data complexity by considering both the bounding box prediction and object classification in active sampling, these techniques are insufficient for active HOI detection as they do not account for the interactions within the images.

(a) (b)

FIGURE 6.1: Illustration of background elements and their impact on active HOI detection. Panels (a) and (b) depict two images, with foreground objects and humans masked in the left counterparts, while the original images are displayed in the right counterparts. Despite foreground items being masked in (a), the contextual clues still allow one to infer that the image illustrates a baseball game. Conversely, in (b), identifying the image's topic is challenging, as the absence of foreground items makes it difficult to recognize the person riding a motorcycle in the air.

One significant challenge is the presence of spurious background information within HOI images, hindering effective active sampling and resulting in suboptimal data selection. Specifically, while background features can contribute to the overall scene context, they can also be detrimental, leading to shortcut learning by the HOI detection model. For example, many previous works have highlighted the detrimental effects of such spurious features [121, 120, 118, 176, 46]. These investigations have demonstrated how misleading background elements can lead the model to learn shortcuts or erroneous associations, rather than focusing on genuine interaction concepts. In certain instances, the model may rely on these spurious features to make predictions, rather than accurately analyzing the human-object pair itself. This phenomenon underscores the importance of robust and effective data selection strategies, especially in the context of active HOI detection. Actively sampling images with significant spurious features can introduce bias and compromise model performance. Therefore, developing an active HOI detection method capable of effectively disregarding spurious background information and prioritizing the sampling of images with informative foreground human-object pairs is crucial.

Fig. 6.1 illustrates background elements and their impact on active HOI detection. For the two images (a) and (b) we provide the original image on the right and the counterpart with foreground items masked. In (a), the baseball bat is not quite visible due to motion blur and the humans are subject to occlusion, whereas in (b), the humans and objects are all clearly presented. Therefore, image (a) contains more image complexity as stated in the first challenge. In this case, image (a) should be sampled due to its difficulty. However, upon examining the masked counterparts of the images, it becomes evident that the background of (a) already provides sufficient information to infer a baseball match, while the background in (b) offers little insight into motorcycle riding. Despite the blurry and occluded objects and humans in image (a), it may not necessarily be more informative for active sampling compared to image (b) due to the spurious features in the background. Therefore, to conduct effective active HOI detection, all elements in the image should be taken into consideration.

To effectively capture complex image scenes while simultaneously addressing the mentioned spurious background features, we propose ActiveHOI, a novel active sampling strategy specially designed for active HOI detection. Specifically, we first employ a foreground masking technique to bypass data with significant spurious features during active sampling. By comparing predictions from original and masked images, ActiveHOI identifies images with substantial background spurious features. The similarity between predictions from the original and masked images is measured with a newly devised comprehensive score, considering multiple image components, bounding box overlappings, and prediction confidences of object and interaction labels, providing a thorough understanding of the HOI image data. Then, we incorporate a diversity-aware sampling strategy to mitigate the selection of similar images during batch active sampling. With the proposed ActiveHOI, we aim to train HOI detection systems with limited annotation budgets while maintaining desirable detection performance. The main contributions of this paper are as follows:

- We investigate the annotation problem in HOI detection, driven by the critical need to efficiently scale up the annotation of rich interaction data within limited budgets, advancing HOI detection research.

- We propose to explore active learning for HOI detection, referred to as ActiveHOI, aiming to maximize the performance of HOI detection models despite limited annotation budgets.
- We conduct extensive experiments to evaluate ActiveHOI and comprehensively analyze the efficacy of each component or technique.

6.2 Related Works

Active Learning. As introduced in previous chapters, existing active learning methods mainly fall in three categories, uncertainty-based [42, 185, 148, 192, 5, 128, 170, 75, 34], disagreement-based [141, 39, 22, 116, 27, 26], and diversity-based [140, 3]. Aside from the conventional active learning methods that are usually evaluated with the image classification task, there are also active learning methods that are specifically designed for other computer vision applications such as semantic segmentation [145, 180, 125], tracking [191], and object detection [20, 190, 175, 117]. Siddiqui et al. [145] proposed to consider active learning for semantic segmentation in a multi-view setting by measuring the inconsistency of predictions across views as the model uncertainty. The work of Yuan et al. [191] introduced a novel active learning method for object tracking in videos. The method integrates the temporal relationship of the moving object in the video sequence for active sampling. To consider both the localization and the classification information in active object detection, Choi et al. [20] combined the aleatoric and epistemic uncertainties. The former measures the intrinsic noise in images and the latter computes the lack of knowledge in the model. Active teacher [117] proposed by Mi et al. dealt with the object detection problem via pseudo-labelling using an exponential mean average teacher. The sampling metric includes both the uncertainty of the box prediction and that of the object classification. Active object detection is closely related to our active HOI detection since it also needs to consider multiple components, box coordinates and object categories, in images during sampling. However, active object detection does not require the prediction of interactions across instances within the image like active HOI detection. To consider such visual understanding in active sampling, we propose ActiveHOI

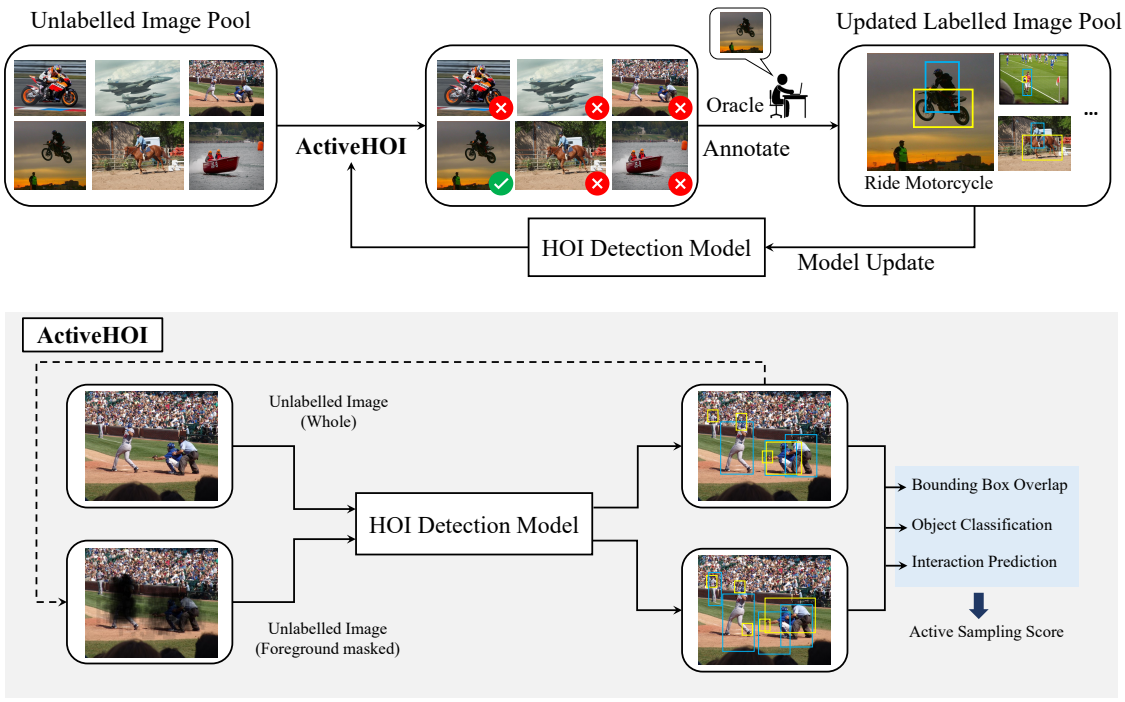


FIGURE 6.2: An overview of the active HOI detection framework is illustrated in the upper part. Our ActiveHOI is shown in the lower part.

which comprehensively fuses uncertainties from multiple components while simultaneously avoiding the spurious features.

Human-Object Interaction Detection. HOI detection was first proposed in [13]. A typical way to solve HOI detection is to use two-stage strategies [13, 131, 47, 53, 98, 205, 106, 43, 113] that separate the instance detection stage and the interaction recognition stage. Li et al. [98] proposed an interactiveness network that measures the interactiveness between each human-object pair. A pose estimator is employed for the network. More recently, one-stage HOI detection methods [152, 15, 74, 208, 132] have been proposed and becoming popular in the field. Tamura et al. [152] proposed QPIC, a query-based HOI detector that simultaneously predicts the bounding boxes, object categories and interaction classes. Qu et al. [132] proposed to improve QPIC by distilling from groundtruth oracle queries instead of using zero vectors. However, the costly annotation has always been an obstacle to HOI detection since complicated visual understanding is required to annotate for such a task. The weakly-supervised learning [92, 162, 158] framework for HOI detection was proposed as a

solution to the problem. This task dismisses the need for bounding box annotations in HOI detection. Unal and Kovashka [158] and Wan et al. [162] proposed to exploit the knowledge from pretrained models. However, this framework still requires extensive annotation of data even with bounding box annotations omitted. Our active HOI detection framework offers an alternative approach to mitigating the annotation costs associated with HOI detection.

6.3 Method

In this section, we begin by offering an overview of the proposed active HOI detection framework. Subsequently, we delve into three key components: the foreground masking strategy, the foreground masked similarity, and the diversity-driven sampling process.

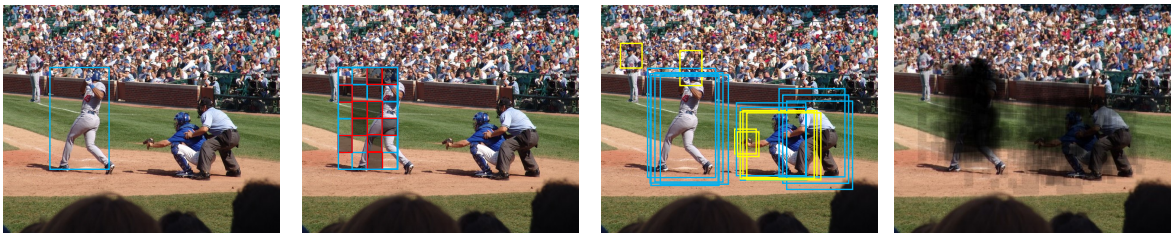
6.3.1 Overview

Fig. 6.2 illustrates the main ActiveHOI framework for active HOI detection. Let $\mathcal{D}_L = \{\mathcal{X}_L, \mathcal{Y}_L\}$ denote a labelled data pool and $\mathcal{D}_U = \{\mathcal{X}_U\}$ denote an unlabelled data pool, where $\mathcal{X}_L$ and $\mathcal{X}_U$ denote the labelled and unlabelled images and $\mathcal{Y}_L$ denotes the HOI annotations for images in $\mathcal{X}_L$. Active HOI detection aims to develop a sampling criterion that can identify a subset of data from $\mathcal{X}_U$ with a fixed budget b to be annotated with an oracle and added to $\mathcal{D}_L$, such that the HOI detection model trained with data in the updated labelled pool can achieve the most desired performance. Given a labelled image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ in $\mathcal{X}_L$ where H is the height and W the width, one or more HOI pairs can exist and be annotated. The annotation of each HOI pair is composed of a normalized human bounding box $\boldsymbol{b}^{(h)} \in [0,1]^4$, a normalized object bounding box $\boldsymbol{b}^{(o)} \in [0,1]^4$, a one-hot object category label $\boldsymbol{c} \in \{0,1\}^{N_{\text{obj}}}$, and the interaction label $\boldsymbol{v} \in \{0,1\}^{N_{\text{act}}}$, where the boxes are normalized based on H and W, N_{obj} denotes the number of possible object categories, and N_{act} denotes the number of possible interaction categories. For simplicity, we adopt the widely used HOI detector QPIC [152] as our baseline model $f(\cdot)$. It is worth noting that the proposed ActiveHOI framework makes no assumptions about HOI detectors and can be easily generalized to other HOI detectors.

6.3.2 Foreground Masking

As depicted in Fig. 6.1, the HOI detection process may inadvertently rely on spurious background features rather than focusing on the relevant humans and objects. This dependence on spurious background elements can introduce noise and inefficiency into the active sampling process. To mitigate this issue, we propose a novel approach involving the masking of foreground boxes. foreground masking involves masking out the detected objects and humans within an image, preserving the surrounding background. By concealing the foreground objects and humans, we can produce HOI detection results that are solely based on the background features. By computing the similarity between predictions from the original image and the masked image, we estimate the impact of background features in the HOI detection process. In the active sampling process, we aim to avoid selecting images that contain spurious backgrounds for HOI detection.

With an image $\boldsymbol{x}$ as input, the output predictions can be generated by the HOI detection model $f(\cdot)$ as $f(\boldsymbol{x}) = \{\hat{\boldsymbol{b}}_i^{(h)}, \hat{\boldsymbol{b}}_i^{(o)}, \hat{\boldsymbol{c}}_i, \hat{\boldsymbol{a}}_i\}_{i=1}^{N_t}$, where $\hat{\boldsymbol{b}}_i^{(h)}$ and $\hat{\boldsymbol{b}}_i^{(o)}$ are predicted normalized bounding boxes, $\hat{\boldsymbol{c}}_i \in [0,1]^4$ is the predicted object category, $\hat{\boldsymbol{a}}_i \in [0,1]^4$ is the predicted interaction label, and N_t stands for the number of proposed HOI pairs. We can mask the foreground objects and humans in $\boldsymbol{x}$ based on the predicted boxes $\hat{\boldsymbol{b}}_i^{(h)}$ and $\hat{\boldsymbol{b}}_i^{(o)}$. The masking strategy is inspired by HaS [147]. For each predicted bounding box $\hat{\boldsymbol{b}}$, we first divide it into patches of size $S \times S$ where S is a predefined positive value. We randomly apply a mask $\boldsymbol{M} \in \{m, 1\}^{3\times H\times W}$ to $\boldsymbol{x}$ with a probability p to mask each patch. Note that the mask value does not differ across channels, therefore we merge the channels and get $\boldsymbol{M} \in \{m, 1\}^{1\times H\times W}$. Also, m stands for the masking coefficient which is a fixed value in the range of $[0, 1)$, where a smaller m indicates a stronger masking effect. p is a fixed probability hyperparameter. The masking process can be summarized as $\tilde{\boldsymbol{x}} = \boldsymbol{x} \prod_{i=1}^{N_t} \boldsymbol{M}_i^{(h)} \boldsymbol{M}_i^{(o)}$, where $\boldsymbol{M}_i^{(h)}$ and $\boldsymbol{M}_i^{(o)}$ stand for the masks generated from the human box and the object box of the i-th HOI pair. While previous image masking works [147, 16, 28] typically utilize a solid mask with $m = 0$, we opt for transparent masks with $m > 0$. The rationale behind this choice stems from the existence of outlier boxes that fail to accurately localize any object or human. Solid masks would treat the outliers and the appropriate predicted boxes equally by masking out the selected patches



(a) (b) (c) (d)

FIGURE 6.3: Illustration of the proposed foreground masking technique. (a) shows one predicted bounding box. (b) shows how box masking is conducted on the predicted box in (a). (c) shows multiple predicted bounding boxes in the image. And (d) shows the result of a complete box masking effect with overlapped masks from multiple predicted boxes.

with the probability p. Recall that we conduct box-level masking and the HOI detection models commonly generate numerous (tens or hundreds of) bounding boxes that overlap with each other within one image. Outliers have a lower chance of overlapping with each other. Therefore, the transparent masking strategy provides more robustness against outlier boxes by masking out more of the overlapped areas. We illustrate the box masking technique in Fig. 6.3 with $m = 0.5$. From the figure, it could be noticed that although each masked patch may not eliminate a significant amount of information from the original image, the multitude of overlapping bounding boxes contributes to creating a compelling masked image.

With the masked image $\tilde{\boldsymbol{x}}$ prepared, we can proceed to compute the similarity between the predictions from the original image $\boldsymbol{x}$ and that from the masked image $\tilde{\boldsymbol{x}}$ to avoid including too much spurious background information in active sampling. The underlying assumption is that if an image yields comparable HOI detection predictions with or without foreground objects and humans masked, it indicates a heavy reliance on background features for the prediction. Consequently, during active sampling, our focus is on selecting images where applying foreground masking leads to lower prediction similarity, indicating less reliance on background features. The computational methodology for prediction similarity will be elucidated in the subsequent subsection.

6.3.3 Foreground Masked Similarity

We evaluate the similarity between the predictions from the original image $\boldsymbol{x}$ and that from the foreground masked image $\tilde{\boldsymbol{x}}$ as follows. Initially, the predictions $f(\boldsymbol{x})$ can be treated as pseudo annotations. Given that the HOI detection model $f(\cdot)$ generates in total $2 \cdot N_t$ human and object boxes in total, redundancy reduction is crucial. Thus, we filter the predictions, retaining only the confident ones. To conduct filtering, we break each predicted HOI pair $\{\hat{\boldsymbol{b}}_i^{(h)}, \hat{\boldsymbol{b}}_i^{(o)}, \hat{\boldsymbol{c}}_i, \hat{\boldsymbol{a}}_i\}$ into N_{act} HOI relationships $\{\hat{\boldsymbol{b}}_i^{(h)}, \hat{\boldsymbol{b}}_i^{(o)}, \hat{\boldsymbol{c}}_i, \hat{a}_{ij}\}_{j=1}^{N_{\text{act}}}$ where $\hat{a}_{ij}$ is a scalar score that stands for the prediction score of interaction j in the i-th predicted HOI pair. We keep only the top-K confident HOI relationships for the pseudo annotations, where K is predefined. The confidence score is computed as $\varphi_{ij} = \max\{\hat{\boldsymbol{c}}_i\}\hat{a}_{ij}$, where $\max\{\hat{\boldsymbol{c}}_i\}$ finds the largest predicted object score. Since the interaction score is split into individuals during this filtering process, there will be $N_t \cdot N_{\text{act}}$ confidence scores computed. If an HOI relationship in the i-th HOI pair is among the top-K confident relationships, we keep its index in a filtered set $\mathcal{F}_i$. Then we need to match the masked bounding boxes, $\tilde{\boldsymbol{b}}^{(o)} \in [0,1]^4$ and $\tilde{\boldsymbol{b}}^{(h)} \in [0,1]^4$, generated via $f(\tilde{\boldsymbol{x}})$, to the pseudo boxes $\hat{\boldsymbol{b}}^{(o)}$ and $\hat{\boldsymbol{b}}^{(h)}$ after filtering. We compute the pairwise generalized IOU of boxes in $f(\tilde{\boldsymbol{x}})$ and boxes in $f(\boldsymbol{x})$ and denote with $\text{GIoU}(\cdot,\cdot)$. The box matching score for the i-th pseudo HOI relationship and the e-th masked HOI relationship is computed as:

$$\sigma_{\text{box}}^{ie} = \theta\big(\text{GIoU}(\hat{\boldsymbol{b}}_i^{(h)}, \tilde{\boldsymbol{b}}_e^{(h)}) \cdot \text{GIoU}(\hat{\boldsymbol{b}}_i^{(o)}, \tilde{\boldsymbol{b}}_e^{(o)})\big), \tag{6.1}$$

where $\theta(q) = q \cdot \mathbb{1}_{q \geq q_0}$ represents a filtering function that eliminates the box matching scores for box pairs that have GIoU scores lower than a threshold q_0. The object similarity score is computed as:

$$\sigma_{\text{obj}}^{ie} = \max\{\hat{\boldsymbol{c}}_i\}\mathbb{1}_{\left\{\arg\max\{\hat{\boldsymbol{c}}_i\} = \arg\max\{\tilde{\boldsymbol{c}}_e\}\right\}}, \tag{6.2}$$

where $\tilde{\boldsymbol{c}}_e$ stands for the category scores computed for the object in the e-th masked HOI pair, and $\hat{\boldsymbol{c}}_i$ is the category scores for the object in the i-th pseudo HOI pair. It represents the object prediction confidence. The similarity computation for the interactions is more complex since each HOI pair can represent multiple interactions. We compute $N_{a_i} = \|\mathcal{F}_i\|$ which is the number of HOI relationships kept for the i-th HOI pair. We then compute the interactions of the top-N_{a_i} interaction scores in $\tilde{\boldsymbol{a}}_e$ and keep the indices in $\tilde{\mathcal{F}}_e$. The interaction similarity

Algorithm 5 Greedy k-Center in ActiveHOI

Data: datasets $\mathcal{D}_L$ and $\mathcal{D}_U$
Hyperparameter: the budget b, the diversity coefficient δ
Model: feature extractor $\phi(\cdot)$ of the HOI detection model
Function: the Euclidean distance function $d(\cdot)$

Find $r = \arg\max_{\boldsymbol{x}_t \in \mathcal{D}_U} \min_{\boldsymbol{x}_k \in \mathcal{D}_L} d(\phi(\boldsymbol{x}_t), \phi(\boldsymbol{x}_k))$
Initialize and update the unlabelled subset $\mathcal{D}_U^s = \{\boldsymbol{x}_r\}$
while $\|\mathcal{D}_U^s\| < \delta \cdot b$ **do**
    Find $r = \arg\max_{\boldsymbol{x}_t \in \mathcal{D}_U / \mathcal{D}_U^s} \min_{\boldsymbol{x}_k \in \mathcal{D}_L \cup \mathcal{D}_U^s} d(\phi(\boldsymbol{x}_t), \phi(\boldsymbol{x}_k))$
    Update the unlabelled subset $\mathcal{D}_U^s = \mathcal{D}_U^s \cup \{\boldsymbol{x}_r\}$
end while
Return the unlabelled subset $\mathcal{D}_U^s$

score is computed as:

$$\sigma_{\text{act}}^{ie} = \frac{1}{N_{a_i}} \sum_{g \in \mathcal{F}_i} \hat{a}_{ig} \mathbb{1}_{\{g \in \tilde{\mathcal{F}}_e\}}. \tag{6.3}$$

To better explain this function, we provide a numerical toy example. Suppose in the i-th pseudo HOI pair, 5 HOI relationships with relatively high prediction confidence are kept in $\mathcal{F}_i$. We then find the interactions in the e-th masked HOI pair that have the top-5 scores. We keep the matched interactions and compute the weighted average with the pseudo HOI interaction scores as weights to include the prediction confidence. The similarity score σ^{ie} for the i-th pseudo HOI pair in $f(\boldsymbol{x})$ and the e-th masked HOI pair in $f(\tilde{\boldsymbol{x}})$ is computed as:

$$\sigma^{ie} = \sigma_{\text{box}}^{ie} \cdot (\sigma_{\text{obj}}^{ie} + \sigma_{\text{act}}^{ie}). \tag{6.4}$$

For each masked HOI pair, we find the largest matching score $\sigma^e = \max_i \{\sigma^{ie}\}$. The image similarity between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ is computed as $\sigma = \sum_{e=1}^{N_t} \sigma^e$.

6.3.4 Diversity-Driven Sampling

To maintain diversity within the selected images, we introduce a diversity-driven sampling process as follows. With the similarity score σ computed for all unlabeled data in $\mathcal{D}_U$, we can directly select images that produce the b smallest similarity scores. However, in the batch active sampling setting, sampling based on per-image informativeness may lead to informative but redundant unlabeled images being selected. To avoid this, we can introduce data diversity during sampling. A simple solution is to sample from a random subset of the

Algorithm 6 ActiveHOI

Data: datasets $\mathcal{D}_L$ and $\mathcal{D}_U$
Hyperparameters: the diversity coefficient δ, the budget b, the number of active cycles N_{active}, the number of proposed HOI pairs N_t
Model: the HOI detection model $f(\cdot)$

for active round 1 to N_{active} **do**
  Train the HOI detection model $f(\cdot)$ with labelled data in $\mathcal{D}_L$
  Find an unlabelled subset $\mathcal{D}_U^s$ of size $\delta \cdot b$ to ensure data diversity.
  for each unlabelled image $\boldsymbol{x} \in \mathcal{D}_U^s$ **do**
    Predict psuedo HOI pairs $f(\boldsymbol{x}) = \{\hat{\boldsymbol{b}}_i^{(h)}, \hat{\boldsymbol{b}}_i^{(o)}, \hat{\boldsymbol{c}}_i, \hat{\boldsymbol{a}}_i\}_{i=1}^{N_t}$ for the original $\boldsymbol{x}$.
    Mask $\boldsymbol{x}$ according to foreground masking in Subsec. 6.3.2 to get $\tilde{\boldsymbol{x}}$
    Predict masked HOI pairs $f(\tilde{\boldsymbol{x}}) = \{\tilde{\boldsymbol{b}}_i^{(h)}, \tilde{\boldsymbol{b}}_i^{(o)}, \tilde{\boldsymbol{c}}_i, \tilde{\boldsymbol{a}}_i\}_{i=1}^{N_t}$ for the masked $\tilde{\boldsymbol{x}}$
    for Masked HOI pair $e = 1$ to N_t **do**
      for Pseudo HOI pair $i = 1$ to N_t **do**
        Find $\mathcal{F}_i$ according to Subsec. 6.3.3
        if $\|\mathcal{F}_i\| > 0$ **then**
          Compute the box matching score σ_{box}^{ie} using Eq. (6.1)
          Compute the object similarity score σ_{obj}^{ie} using Eq. (6.2)
          Compute the interaction similarity σ_{act}^{ie} using Eq. (6.3)
          Compute the overall similarity score σ^{ie} using Eq. (6.4)
        end if
      end for
      Find the similarity score $\sigma^e = \max_i\{\sigma^{ie}\}$ between masked pair e and the most matched pseudo pair.
    end for
    Compute the image similarity $\sigma = \sum_{e=1}^{N_t} \sigma^e$ between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ for sampling
  end for
  Send b unlabelled images with the smallest σ to the oracle for annotation
  Remove the annotated data from $\mathcal{D}_U$ and add to $\mathcal{D}_L$
end for
Return updated $\mathcal{D}_L$ and $\mathcal{D}_U$

unlabelled pool [5, 185], denoted as $\mathcal{D}_U^s$. The size of the unlabelled subset is $\delta \cdot b$, where δ represents an integer-valued diversity coefficient and b is the sampling budget as mentioned above. Alternatively, we can also find the unlabelled subset $\mathcal{D}_U^s$ using a greedy k-center algorithm [**dasgupta2004analysis**, 140] to improve the sampling diversity. The steps are summarized in Alg 5. The complete sampling process is summarized in Alg. 6. The image similarity score not only shows the latent spurious features carried in an image but also integrates the uncertainty of predictions.

6.4 Experiments

6.4.1 Implementation Details

We conducted comprehensive experiments on the HICO-DET dataset [13] and the V-COCO dataset [52, 101], following the standard evaluation protocol. HICO-DET comprises $38,118$ images for training and $9,658$ images for testing. Each image is annotated with labels from $N_{\text{obj}} = 80$ object classes and $N_{\text{act}} = 117$ interaction classes. For the active sampling setting, we initially start with $1,000$ labelled images. The budget b is also set to $1,000$. We repeat the active sampling process until the labelled pool reaches $5,000$ ($N_{\text{active}} = 5$). V-COCO comprises $5,400$ images for training and $4,946$ images for testing. Each image is annotated with labels from $N_{\text{obj}} = 80$ object classes and $N_{\text{act}} = 26$ interaction classes. For the active sampling setting, we initially start with 250 labelled images. The budget b is also set to 250. We repeat the active sampling process until the labelled pool reaches $1,250$ ($N_{\text{active}} = 5$).

QPIC [152] is adopted for the HOI detection model $f(\cdot)$. The model undergoes pretraining using images sourced from the COCO dataset [101]. The pretraining is conducted separately for HICO-DET and V-COCO, ensuring that there is no overlap between the images used for pretraining and those used for HOI detection. We use ResNet-50 [56] as the backbone model. The model is trained for 100 epochs with batch size 8 on 4 NVIDIA Tesla V100 GPUs. The number of predicted HOI pairs follows QPIC to be $N_t = 100$. The hyperparameters for box masking are set to $m = 0.9$, $p = 0.1$, and $S = 20$. To filter out the less confident HOI relationships, we set $K = 100$ and $q_0 = 0.5$. For diversity in sampling, we set the diversity coefficient as $\delta = 10$.

6.4.2 Results

We compare our ActiveHOI with 5 other active sampling strategies and show the results in Table 6.1. The training process remains the same for all strategies. (I) Random sampling is just randomly sampling from the unlabelled data pool for annotation. (II) Margin [5], Entropy [142], and LL4AL [185] fall into the uncertainty-based active learning category.

TABLE 6.1: Comparing results from ActiveHOI with the results from previous active learning baselines. $N_L = \|\mathcal{D}_L\|$ represents the number of data points in the labelled pool. The best result numbers are in **bold** and the second best result numbers are underlined.

HICO-DET					
Methods	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
Random	11.33	13.32	15.66	16.82	18.21
Margin [5]		13.02	14.60	15.71	17.18
Entropy [142]		12.70	14.89	16.59	16.93
LL4AL [185]		13.16	16.03	17.19	18.42
CoreSet [140]		14.01	16.38	17.33	18.52
ActiveHOI		13.95	16.42	17.75	19.21
ActiveHOI-diverse		**14.96**	**16.92**	**18.26**	**20.17**
V-COCO					
Methods	$N_L = 250$	$N_L = 500$	$N_L = 750$	$N_L = 1,000$	$N_L = 1,250$
Random	15.15	18.48	20.47	23.69	30.18
Margin [5]		20.42	21.28	26.99	33.55
Entropy [142]		20.97	22.12	27.27	34.61
LL4AL [185]		16.37	17.46	24.48	30.36
CoreSet [140]		16.45	18.4	22.50	30.72
ActiveHOI		**22.74**	**24.60**	**28.73**	**35.14**
ActiveHOI-diverse		16.33	20.11	26.41	32.3

Margin means to actively sample data that generate small margin scores. The margin score for one object box in each image is computed as the highest object prediction score subtracting the second-highest one. The margin scores for all object boxes are averaged to become the image margin score for sampling. Smaller margin scores indicate lower confidence and higher uncertainty in predictions. The entropy score is similarly averaged but higher values are favoured instead. Note that the predicted interaction scores cannot be used to compute uncertainty since one pair of a human and an object can have multiple interactions involved. To sample with entropy scores, we select images that produce high entropy scores which indicate high uncertainties. LL4AL is a special type of uncertainty-based active learning method that predicts the training loss of each image with an extra module. (III) CoreSet [140] is a diversity-based active learning method that selects data with features spanning over the space. For our approach, we present two sets of results, 'ActiveHOI' denotes the outcomes obtained by randomly selecting the unlabelled subset $\mathcal{D}_U^s$, while 'ActiveHOI-diverse' signifies

the results achieved by incorporating additional diversity through sampling the unlabelled subset $\mathcal{D}_U^s$ using a greedy k-center algorithm. Further details can be found in Subsec. 6.3.4.

According to the HICO-DET results presented in Table 6.1, our ActiveHOI-diverse outperforms all other baselines, including our own ActiveHOI, except for CoreSet with $N_L = 2,000$. Our ActiveHOI falls short only in comparison to CoreSet at this specific label pool size. The reason for this discrepancy may lie in the relative importance of data diversity when the labeled data pool is smaller. It is noteworthy that conventional uncertainty-based methods, such as margin and entropy scores, fail to produce satisfactory results, especially when compared to random sampling. This inadequacy could stem from their inability to consider data diversity effectively. Margin scores and entropy scores may tend to favor certain objects without adequately considering similarities among images. Moreover, solely relying on object prediction scores may not fully capture the informativeness of complex images containing humans, interactions, and surrounding scenes. Consequently, margin scores and entropy scores might select data that lack both individual informativeness and collective diversity. In contrast, LL4AL achieves superior results by incorporating multiple components of the image in the loss prediction process.

The results for V-COCO are also depicted in Table 6.1. In contrast to the HICO-DET findings, ActiveHOI outperforms all other methods, while ActiveHOI-diverse does not perform as well. Interestingly, conventional uncertainty-based methods yield competitive results in this context. Conversely, LL4AL, CoreSet, and our ActiveHOI-diverse exhibit suboptimal performance. These discrepancies may arise from the differing characteristics of the datasets. HICO-DET boasts a significantly larger training data size compared to V-COCO (38k vs. 5k), indicating that V-COCO places less emphasis on data diversity during active sampling. Moreover, LL4AL, CoreSet, and our ActiveHOI-diverse are all heavily reliant on the feature extractor of the HICO detection model. LL4AL predicts loss using the extracted features, while CoreSet and ActiveHOI-diverse compute Euclidean distances for data diversity based on the extracted image features. This suggests that the feature extractor trained with V-COCO may not perform as effectively as that trained with HICO-DET. In other words, classifiers for V-COCO play a more crucial role in HOI detection, possibly because the object classifier for

V-COCO is easier to train with a smaller label space. Additionally, the smaller pretraining set for V-COCO, resulting from the exclusion of overlapping images, could also contribute to these disparities.

6.4.3 Ablation Studies

To evaluate our proposed methods, we conduct ablation studies on HICO-DET to validate the design of each component. To better analyze each module, we conduct all but one ablation experiment with ActiveHOI. We analyze the diversity coefficient δ with ActiveHOI-diverse.

TABLE 6.2: Ablation studies on ActiveHOI by using different masking strategies.

Masking Strategy	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
Random full	11.33	13.90	15.58	16.46	17.49
Random box		13.23	15.57	17.13	18.47
Foreground box (ours)		13.95	16.42	17.75	19.21

Masking strategy. ActiveHOI adopts a unique foreground masking technique which masks repeatedly according to the pseudo bounding boxes and achieves satisfying experiment results. However, could the performance gain be coming from the noise caused by masking instead of from the deliberate removal of foreground instances? Hence we design this experiment which replaces the foreground masking strategy with two random masking strategies. 'Random full' represents the masking strategy where we mask randomly on the whole image instead of within boxes. 'Random box' stands for another masking strategy where we mask within randomly located boxes. We still use the sizes of the pseudo boxes to retain the intensity of masking. But we randomly change the location of the masks so that they do not deliberately mask the foreground instances. The results are shown in Table 6.2. Our foreground masking strategy outperforms the two random masking strategies with a clear margin. Furthermore, random box masking outperforms random full masking, possibly because of the retained masking intensity, demonstrating the reasonability of our masking strategy.

Components in σ computation. We analyze the components in computing the similarity score σ and show the results in Table 6.3. We progressively remove σ_{box}, σ_{obj}, and σ_{act}

TABLE 6.3: Ablation studies on ActiveHOI by removing components in the similarity score computation.

σ_{box}	σ_{obj}	σ_{act}	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
	✓		11.33	13.41	15.56	16.92	18.45
		✓		13.47	15.77	17.10	18.56
	✓	✓		13.52	15.97	17.03	18.89
✓	✓	✓		13.95	16.42	17.75	19.21

to assess their individual contributions. Starting from the bottom of the table and moving upwards, removing σ_{box} means that we no longer favour larger overlaps while matching the masked HOI pairs to the pseudo HOI pairs. Additionally, box overlaps are disregarded during active sampling. Results show that σ_{box} is indispensable. Without σ_{obj} and σ_{act} respectively, the performance also dropped. Removing σ_{act} incurred a more significant degradation in performance compared to removing σ_{obj}, possibly because σ_{act} computes from multiple predictions of interactions and contains more useful information than σ_{obj}. This experimentation underscores the necessity of each component in the design of σ computation.

TABLE 6.4: Ablation studies on ActiveHOI by using a simple similarity computation.

Masking Strategy	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
Feature-wise	11.33	13.52	15.78	17.01	18.57
ActiveHOI (ours)		13.95	16.42	17.75	19.21

Similarity computation. We also experiment by replacing the similarity σ computation with a simpler similarity computation where we simply compute the Euclidean distance between features extracted from the original image and the masked image. The results are shown in Table 6.4. Our ActiveHOI with a comprehensive similarity outperforms the simple feature-wise similarity. The results further verify that the similarity computation should include multiple components, bounding boxes, object labels, and interaction labels, in images. Extracted image features do not carry enough information for effective active sampling.

TABLE 6.5: Ablation studies on the diversity coefficient δ for ActiveHOI-diverse.

Diversity	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
$\delta = 5$	11.33	15.01	17.06	18.89	19.71
$\delta = 10$		14.96	16.92	18.26	20.17
$\delta = 15$		14.11	16.49	18.12	19.70

TABLE 6.6: Results on HICO-DET from using different values of the masking coefficient m and the masking probability p. ActiveHOI is used for the experiments.

m	p	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
0.1	0.1	11.33	12.68	15.32	17.69	17.78
0.1	0.5		12.81	16.02	16.99	18.45
0.1	0.9		13.34	15.22	16.60	18.75
0.5	0.1		12.94	15.18	17.31	18.88
0.5	0.5		12.86	15.57	17.11	18.30
0.5	0.9		13.18	15.73	16.76	18.44
0.9	0.1		13.95	16.42	17.75	19.21
0.9	0.5		13.13	16.26	17.34	18.95
0.9	0.9		12.59	15.37	16.14	18.22

Diversity coefficient. We analyze ActiveHOI-diverse by conducting experiments with various values of the diversity coefficient δ. Our findings indicate that using $\delta = 5$ and $\delta = 10$ yields better results compared to using $\delta = 15$. A smaller δ value imposes a stronger constraint on data diversity during sampling, while larger δ values ease this constraint. There is a trade-off between the sampled data diversity and the individual informativeness of each sampled data point. Given its superior performance with $N_L = 5,000$, we opt for $\delta = 10$.

Masking Intensity. While conducting our foreground masking, the values of the masking coefficient m, how much information in each patch is kept, and the masking probability p, the probability of a patch within a bounding box being masked, should be carefully chosen. With the value of m increasing, more information is kept after each masking. With the value of p increasing, more patches are masked for each bounding box. After conducting the experiments with 9 pairs of m and p values, with results shown in Table 6.6, we found that $m = 0.9$ and $p = 0.1$ can produce the most satisfying results. $m = 0.9$ and $p = 0.5$ ranks the second in performance. The underlying reason could be that using smaller values of m might reduce the robustness of the method. Outlier boxes could affect the masking when m adopts smaller values. For example, to reduce the information in a pixel to 10% of the original one, $m = 0.1$ takes only one mask, $m = 0.5$ takes 4 masks, and $m = 0.9$ takes 22 masks. The

TABLE 6.7: Results on HICO-DET from using different values of the masking grid size S. ActiveHOI is used for the experiments.

S	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
10	11.33	13.87	16.29	17.91	19.02
20		13.95	16.42	17.75	19.21
30		13.67	16.34	17.23	18.98
40		13.25	15.83	16.89	18.74
50		13.52	15.76	17.08	18.81

TABLE 6.8: Results on HICO-DET from using different values of K for keeping the top-K pseudo HOI relationships. ActiveHOI is used for the experiments.

K	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
50	11.33	14.03	16.10	17.21	18.79
100		13.95	16.42	17.75	19.21
200		13.47	16.24	17.51	19.03
∞		13.11	15.98	17.22	18.45

value of p also controls the intensity of masking, similar effects apply. Therefore, we choose $m = 0.9$ and $p = 0.1$ for our ActiveHOI.

Masking Grid Size. Another topic to be studied is the grid size used for masking. The grid size also affects the masking intensity. A larger grid size implies more intensity in masking because larger grids make it harder to discern the masked patches from neighbouring patches, despite maintaining the overall masked area constant regardless of the grid size. We conduct experiments for 5 different choices of the grid size S and show the results in Table 6.7. Generally, using smaller grid sizes leads to better experiment results. Consequently, based on these results, we select $S = 20$ for our ActiveHOI based on the results.

Top-K Confident for Filtering. We also investigate the impact of selecting different values for K when constructing each filtered set $\mathcal{F}$. Specifically, we experiment with top-50, top-100, and top-200 confident pseudo HOI relationships in filtering. Additionally, we explore the scenario with no filtering, denoted by $K = \infty$, as shown in Table 6.8. The results show

TABLE 6.9: Results on HICO-DET from using different filtering techniques. ActiveHOI is used for the experiments.

Filtering	$N_L = 1,000$	$N_L = 2,000$	$N_L = 3,000$	$N_L = 4,000$	$N_L = 5,000$
NMS	11.33	13.34	15.81	17.23	18.36
Confidence		13.95	16.42	17.75	19.21

that smaller K values can lead to better performance when the labelled pool is relatively small, and larger K values perform better with larger labelled pools. It is not difficult to infer that when the model is not powerful enough, the confidence in prediction is low, hence a smaller K is favoured. Similarly, with a more powerful model, a larger K is preferred. Furthermore, eliminating the filtering process does not yield satisfactory results, likely due to the introduction of noise from predictions with low confidence. Based on these insights, we select $K = 100$ for our ActiveHOI method.

Non-Maximum Suppression for Filtering. In addition to our confidence-based filtering technique, we can also employ non-maximum suppression (NMS) to filter the pseudo HOI pairs. To conduct NMS, we iteratively remove HOI pairs that overlap with others and exhibit lower confidence in prediction. Overlapping is determined by a generalized IoU threshold of greater than 0.5 for both the human and object boxes compared to another pair. And the confidence is represented by the largest object classification score for each object box. The results are shown in Table 6.9. Our confidence-based filtering outperforms NMS. The possible reason is that our confidence score considers both the interaction prediction score and the object classification score.

6.5 Future Work

This work opens up several promising avenues for future research. One direction involves extending the proposed method to video-based HOI detection, where temporal dynamics play a crucial role. Incorporating temporal factors into the active sampling process could enhance the model's performance in capturing dynamic interactions over time while still controlling annotation costs. Additionally, the integration of our active HOI detection approach with

weakly-supervised learning techniques presents another promising area for exploration. By leveraging weakly-supervised annotations alongside active learning, we could further reduce annotation costs while maintaining satisfying performance. This integration between active and weakly-supervised learning methodologies could offer substantial benefits in real-world applications of HOI detection.

6.6 Conclusion

In this work, we delved into the active Human-Object Interaction (HOI) detection problem, which endeavours to enhance HOI detection models by selecting informative unlabeled data for annotation. This task involves actively annotating informative unlabelled data to improve localizing and recognizing human-object pairs within images, presenting significant challenges due to the diverse components and complex scenes depicted in the imagery. To address these challenges, we employed a foreground masking technique to avoid spurious features alongside a comprehensive similarity computation approach for sampling. Our experimental evaluation was carried out on two widely recognized HOI detection benchmarks, where we compared the performance of our proposed approach against state-of-the-art active learning methods. Through extensive experimentation and analyses, we demonstrated the effectiveness and superiority of our method. Our study contributes novel insights and methodologies to active HOI detection, offering practical solutions to address the challenges posed by complex images.

CHAPTER 7

Conclusion

7.1 Summary

The thesis delved into the multifaceted challenges posed by data insufficiency in various learning tasks within the realm of deep learning. Through a comprehensive exploration of attribute-based person search, privacy-preserving active learning methodologies, and active HOI detection, we have contributed novel insights and methodologies to address these pressing issues.

The thesis first proposed SAL for attribute-based person search. SAL preserves finely-grained discriminative information and generates more reliable synthetic features to augment the common embedding space. SAL effectively synthesizes visual representations of unseen classes, making it more robust in zero-shot retrieval scenarios commonly encountered in real-world person search tasks. Additionally, SAM, an extension of SAL, was proposed to enrich the sparse feature space and to bridge the semantic gap between modalities.

Furthermore, the thesis introduced the Peer Study Learning (PSL) framework for responsible active learning, which simultaneously prioritizes data privacy preservation and model stability. By employing a teacher-student architecture and a novel learning paradigm, PSL achieves superior results over existing active learning methods across both sensitive protection settings and standard benchmarks. The thesis also provided another aspect in addressing efficient active sampling together with data privacy preserving. We proposed the Knowledge-Aware Federated Active Learning (KAFAL) method to address the federated active learning problem. KAFAL effectively mitigates discrepancies arising from non-IID data distributions among

client-server models, thus ensuring the robustness and performance of global models in decentralized learning scenarios.

The thesis also explored active learning in special applications. We introduced the active Human-Object Interaction (HOI) detection problem, which endeavours to enhance HOI detection models by selecting informative unlabeled data for annotation. This task involves actively annotating informative unlabelled data to improve localizing and recognizing human-object pairs within images, presenting significant challenges due to the diverse components and complex scenes depicted in the imagery. To address these challenges, we employed a foreground masking technique to avoid spurious features alongside a comprehensive similarity computation approach for sampling.

In conclusion, the thesis explored and addressed the data insufficiency problem in deep learning from mainly two streams. One is via data augmentation to tackle the unseen possible data. Another is to optimally use the limited annotation budgets to tackle the insufficiency of data labels.

7.2 Future Studies

The issue of data insufficiency in deep learning has long been a focal point of study, and while this thesis has addressed a subset of this vast topic, there remains ample room for future exploration.

With the emergence of high-quality image generation models like those discussed in recent literature [136], there arises an intriguing shift in perspective. Rather than solely focusing on how to generate unseen data, it may be prudent to consider what to generate. This opens the door to innovative approaches that combine active learning techniques with the generation of novel data. Future investigations could delve into the challenge of generating informative data with specific annotations, presenting an exciting avenue for further research.

Moreover, the evaluation of active learning methods poses a persistent challenge within the research community. Presently, researchers often devise their own data splits for training,

sampling, and evaluation, leading to inconsistencies and inefficiencies in comparison across studies. Addressing this issue is crucial for advancing the field in a more streamlined and environmentally sustainable manner. Future studies could explore novel approaches to standardize and streamline the evaluation process, facilitating fair and efficient comparisons between different methodologies. This would not only save time and resources but also foster greater progress within the research community.

Another promising avenue for future research lies in establishing upper bounds for the performance of active learning across different benchmarks. While it's widely acknowledged that active learning provides benefits over passive learning strategies, the extent to which active learning can improve performance varies depending on the dataset characteristics. Future work could seek to determine the maximum potential improvement that active learning can offer for datasets with varying distributions. For instance, in datasets with more uniformly distributed samples, the utility of active learning may diminish compared to datasets with highly imbalanced distributions. The label space of different datasets may also affect the performance upper bound, even with the same images. Finding a way to quantify the potential for active learning in different datasets could be helpful. This understanding can inform the development of more tailored and effective active learning strategies, particularly for datasets where the benefits of active learning are less pronounced.

Bibliography

[1] Jin-Hyun Ahn, Kyungsang Kim, Jeongwan Koh and Quanzheng Li. 'Federated Active Learning (F-AL): an Efficient Annotation Strategy for Federated Learning'. In: *arXiv preprint arXiv:2202.00195* (2022).

[2] Galen Andrew, Raman Arora, Jeff Bilmes and Karen Livescu. 'Deep canonical correlation analysis'. In: *ICML*. 2013.

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford and Alekh Agarwal. 'Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds.' In: *ICLR*. 2020.

[4] Elnaz Barshan and Paul Fieguth. 'Stage-wise training: An improved feature learning strategy for deep models'. In: *Feature Extraction: Modern Questions and Challenges*. 2015, pp. 49–59.

[5] William H Beluch, Tim Genewein, Andreas Nürnberger and Jan M Köhler. 'The power of ensembles for active learning in image classification'. In: *CVPR*. 2018.

[6] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang and Colin Raffel. 'Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring'. In: *arXiv preprint arXiv:1911.09785* (2019).

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver and Colin Raffel. 'Mixmatch: A holistic approach to semi-supervised learning'. In: *arXiv preprint arXiv:1905.02249* (2019).

[8] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan and Ting Yao. 'Exploring object relation in mean teacher for cross-domain detection'. In: *CVPR*. 2019.

[9] Yu-Tong Cao, Jingya Wang and Dacheng Tao. 'Symbiotic Adversarial Learning for Attribute-based Person Search'. In: *European Conference on Computer Vision*. Springer. 2020, pp. 230–247.

[10] Razvan Caramalau, Binod Bhattarai and Tae-Kyun Kim. 'Sequential graph convolutional network for active learning'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9583–9592.

[11] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh and Christopher J Pal. 'Reinforced active learning for image segmentation'. In: *arXiv preprint arXiv:2002.06583* (2020).

[12] Sergio Casas, Abbas Sadat and Raquel Urtasun. 'Mp3: A unified model to map, perceive, predict and plan'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14403–14412.

[13] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng and Jia Deng. 'Learning to detect human-object interactions'. In: *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE. 2018, pp. 381–389.

[14] Hong-You Chen and Wei-Lun Chao. 'On Bridging Generic and Personalized Federated Learning for Image Classification'. In: *International Conference on Learning Representations*. 2021.

[15] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang and Chen Qian. 'Reformulating hoi detection as adaptive set prediction'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9004–9013.

[16] Pengguang Chen, Shu Liu, Hengshuang Zhao and Jiaya Jia. 'Gridmask data augmentation'. In: *arXiv preprint arXiv:2001.04086* (2020).

[17] Xu Chen and Brett Wujek. 'Autodal: Distributed active learning with automatic hyperparameter selection'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 3537–3544.

[18] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng and Jian-Huang Lai. 'Person re-identification by camera correlation aware feature augmentation'. In: *IEEE TPAMI* 40.2 (2018).

[19] Kashyap Chitta, Aditya Prakash and Andreas Geiger. 'Neat: Neural attention fields for end-to-end autonomous driving'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15793–15803.

[20] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet and Jose M Alvarez. 'Active learning for deep object detection via probabilistic modeling'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10264–10273.

[21] LI Chongxuan, Taufik Xu, Jun Zhu and Bo Zhang. 'Triple generative adversarial nets'. In: *NIPS*. 2017.

[22] David Cohn, Les Atlas and Richard Ladner. 'Improving generalization with active learning'. In: *Machine learning* 15.2 (1994), pp. 201–221.

[23] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec and Matei Zaharia. 'Selection via Proxy: Efficient Data Selection for Deep Learning'. In: *ICLR*. 2019.

[24] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth and Bernt Schiele. 'The cityscapes dataset for semantic urban scene understanding'. In: *CVPR*. 2016.

[25] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri and Ningshan Zhang. 'Region-based active learning'. In: *22nd AISTATS*. 2019.

[26] Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, Ningshan Zhang and Claudio Gentile. 'Active learning with disagreement graphs'. In: *ICML*. 2019.

[27] Ido Dagan and Sean P Engelson. 'Committee-based sampling for training probabilistic classifiers'. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 150–157.

[28] Terrance DeVries and Graham W Taylor. 'Improved regularization of convolutional neural networks with cutout'. In: *arXiv preprint arXiv:1708.04552* (2017).

[29] Yubin Deng, Ping Luo, Chen Change Loy and Xiaoou Tang. 'Learning to recognize pedestrian attribute'. In: *arXiv:1501.00901* (2015).

[30] Yubin Deng, Ping Luo, Chen Change Loy and Xiaoou Tang. 'Pedestrian attribute recognition at far distance'. In: *ACM MM*. ACM. 2014.

[31] Mandar Dixit, Roland Kwitt, Marc Niethammer and Nuno Vasconcelos. 'Aga: Attribute-guided augmentation'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7455–7463.

[32] Qi Dong, Shaogang Gong and Xiatian Zhu. 'Person Search by Text Attribute Query As Zero-Shot Learning'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 3652–3661.

[33] Xuanyi Dong and Yi Yang. 'Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection'. In: *ICCV*. 2019.

[34] Sayna Ebrahimi, William Gan, Dian Chen, Giscard Biamby, Kamyar Salahi, Michael Laielli, Shizhan Zhu and Trevor Darrell. 'Minimax active learning'. In: *arXiv preprint arXiv:2012.10467* (2020).

[35] Aviv Eisenschtat and Lior Wolf. 'Linking image and text with 2-way nets'. In: *CVPR*. 2017.

[36] Rafael Felix, Vijay BG Kumar, Ian Reid and Gustavo Carneiro. 'Multi-modal cycle-consistent generalized zero-shot learning'. In: *ECCV*. 2018.

[37] Fangxiang Feng, Xiaojie Wang and Ruifan Li. 'Cross-modal retrieval with correspondence autoencoder'. In: *ACM MM*. ACM. 2014.

[38] Matt Fredrikson, Somesh Jha and Thomas Ristenpart. 'Model inversion attacks that exploit confidence information and basic countermeasures'. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333.

[39] Yoav Freund, H Sebastian Seung, Eli Shamir and Naftali Tishby. 'Information, prediction, and query by committee'. In: *NIPS*. 1993.

[40] Bo Fu, Zhangjie Cao, Jianmin Wang and Mingsheng Long. 'Transferable Query Selection for Active Domain Adaptation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7272–7281.

[41] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti and Anima Anandkumar. 'Born again neural networks'. In: *arXiv preprint arXiv:1805.04770* (2018).

[42] Yarin Gal, Riashat Islam and Zoubin Ghahramani. 'Deep Bayesian Active Learning with Image Data'. In: *ICML*. 2017.

[43] Chen Gao, Jiarui Xu, Yuliang Zou and Jia-Bin Huang. 'Drg: Dual relation graph for human-object interaction detection'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer. 2020, pp. 696–712.

[44] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang and Shih-Fu Chang. 'Low-shot learning via covariance-preserving adversarial augmentation networks'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 975–985.

[45] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis and Tomas Pfister. 'Consistency-based semi-supervised active learning: Towards minimizing labeling cost'. In: *European Conference on Computer Vision*. Springer. 2020, pp. 510–526.

[46] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge and Felix A Wichmann. 'Shortcut learning in deep neural networks'. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[47] Georgia Gkioxari, Ross Girshick, Piotr Dollár and Kaiming He. 'Detecting and recognizing human-object interactions'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8359–8367.

[48] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu and Anuj Kumar. 'Active federated learning'. In: *arXiv preprint arXiv:1909.12641* (2019).

[49] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann and Arun Innanje. 'Ensemble Attention Distillation for Privacy-Preserving Federated Learning'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15076–15086.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. 'Generative adversarial nets'. In: *NIPS*. 2014.

[51] Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi and Sotaro Tsukizawa. 'Deep active learning for biased datasets via fisher kernel self-supervision'. In: *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 9041–9049.

[52] Saurabh Gupta and Jitendra Malik. 'Visual semantic role labeling'. In: *arXiv preprint arXiv:1505.04474* (2015).

[53] Tanmay Gupta, Alexander Schwing and Derek Hoiem. 'No-frills human-object interaction detection: Factorization, layout encodings, and training techniques'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9677–9685.

[54] David R Hardoon, Sandor Szedmak and John Shawe-Taylor. 'Canonical correlation analysis: An overview with application to learning methods'. In: *Neural computation* 16.12 (2004).

[55] Bharath Hariharan and Ross Girshick. 'Low-shot visual recognition by shrinking and hallucinating features'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3018–3027.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)*. 2015.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 'Deep residual learning for image recognition'. In: *CVPR*. 2016.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 'Identity mappings in deep residual networks'. In: *ECCV*. Springer. 2016.

[59] Alexander Hermans, Lucas Beyer and Bastian Leibe. 'In defense of the triplet loss for person re-identification'. In: *arXiv:1703.07737* (2017).

[60] Geoffrey Hinton, Oriol Vinyals and Jeff Dean. 'Distilling the knowledge in a neural network'. In: *arXiv preprint arXiv:1503.02531* (2015).

[61] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros and Trevor Darrell. 'Cycada: Cycle-consistent adversarial domain adaptation'. In: *arXiv:1711.03213* (2017).

[62] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan et al. 'Searching for mobilenetv3'. In: *ICCV*. 2019.

[63] Kevin Hsieh, Amar Phanishayee, Onur Mutlu and Phillip Gibbons. 'The non-iid data quagmire of decentralized machine learning'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4387–4398.

[64] Tzu-Ming Harry Hsu, Hang Qi and Matthew Brown. 'Federated visual classification with real-world data distribution'. In: *European Conference on Computer Vision*. Springer. 2020, pp. 76–92.

[65] Tzu-Ming Harry Hsu, Hang Qi and Matthew Brown. 'Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification'. In: *arXiv preprint arXiv:1909.06335* (2019).

[66] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan and Dejing Dou. 'Semi-supervised active learning with temporal output discrepancy'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3447–3456.

[67] Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li and Qi Tian. 'ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1235–1244.

[68] David Isele, Reza Rahimi, Akansel Cosgun, Kaushik Subramanian and Kikuo Fujimura. 'Navigating occluded intersections with autonomous vehicles using deep reinforcement learning'. In: *2018 ICRA*. IEEE. 2018.

[69] Emad Sami Jaha and Mark S Nixon. 'Soft biometrics for subject identification using clothing attributes'. In: *IJCB*. IEEE. 2014.

[70] Harold Jeffreys. 'Theory of Probability'. In: (1939).

[71] Boseung Jeong, Jicheol Park and Suha Kwak. 'Asmr: Learning attribute-based person search with adaptive semantic margin regularizer'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12016–12025.

[72] Wonyong Jeong, Jaehong Yoon, Eunho Yang and Sung Ju Hwang. 'Federated semi-supervised learning with inter-client consistency & disjoint learning'. In: *arXiv preprint arXiv:2006.12097* (2020).

[73] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu and Dilip Krishnan. 'Supervised contrastive learning'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18661–18673.

[74] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim and Hyunwoo J Kim. 'Hotr: End-to-end human-object interaction detection with transformers'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 74–83.

[75] Kwanyoung Kim, Dongwon Park, Kwang In Kim and Se Young Chun. 'Task-aware variational adversarial active learning'. In: *CVPR*. 2021.

[76] SangMook Kim, SangMin Bae, Se-Young Yun and Hwanjun Song. 'LG-FAL: Federated Active Learning Strategy using Local and Global Models'. In: ().

[77] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee and Jiwon Kim. 'Learning to discover cross-domain relations with generative adversarial networks'. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1857–1865.

[78] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization'. In: *arXiv:1412.6980* (2014).

[79] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage and Peter Richtárik. 'Federated optimization: Distributed machine learning for on-device intelligence'. In: *arXiv preprint arXiv:1610.02527* (2016).

[80] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009.

[81] Alex Krizhevsky, Geoffrey Hinton et al. 'Learning multiple layers of features from tiny images'. In: (2009).

[82] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. 'Imagenet classification with deep convolutional neural networks'. In: *Advances in neural information processing systems* 25 (2012).

[83] Solomon Kullback and Richard A Leibler. 'On information and sufficiency'. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[84] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra and Piyush Rai. 'Generalized zero-shot learning via synthesized examples'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4281–4289.

[85] Ryan Layne, Timothy M Hospedales and Shaogang Gong. 'Attributes-based re-identification'. In: *Person Re-Identification*. Springer, 2014.

[86] Ryan Layne, Timothy M Hospedales and Shaogang Gong. 'Towards person identification and re-identification with attributes'. In: *ECCV*. Springer. 2012.

[87] Ryan Layne, Timothy M Hospedales, Shaogang Gong and Q Mary. 'Person re-identification by attributes.' In: *BMVC*. 2012.

[88] Ya Le and Xuan Yang. 'Tiny imagenet visual recognition challenge'. In: *CS 231N* 7.7 (2015), p. 3.

[89] Yann LeCun, Corinna Cortes and CJ Burges. 'MNIST handwritten digit database'. In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[90] Dangwei Li, Xiaotang Chen and Kaiqi Huang. 'Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios'. In: *IAPR ACPR*. IEEE. 2015.

[91] Liulei Li, Jianan Wei, Wenguan Wang and Yi Yang. 'Neural-logic human-object interaction detection'. In: *Advances in Neural Information Processing Systems* 36 (2024).

[92] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic and Bryan Russell. 'Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1845–1855.

[93] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang and Xiaogang Wang. 'Identity-aware textual-visual matching with latent co-attention'. In: *ICCV*. 2017.

[94] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar and Virginia Smith. 'Federated optimization in heterogeneous networks'. In: *Proceedings of Machine Learning and Systems* 2 (2020), pp. 429–450.

[95] Wei Li, Rui Zhao, Tong Xiao and Xiaogang Wang. 'Deepreid: Deep filter pairing neural network for person re-identification'. In: *CVPR*. 2014.

[96] Wei Li, Xiatian Zhu and Shaogang Gong. 'Person re-identification by deep joint learning of multi-loss classification'. In: *arXiv:1705.04724* (2017).

[97] Xin Li and Yuhong Guo. 'Adaptive active learning for image classification'. In: *CVPR*. 2013.

[98] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang and Cewu Lu. 'Transferable interactiveness knowledge for human-object interaction detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3585–3594.

[99] Liang Lin, Keze Wang, Deyu Meng, Wangmeng Zuo and Lei Zhang. 'Active self-paced learning for cost-effective and progressive face identification'. In: *TPAMI* 40.1 (2017), pp. 7–19.

[100] Tao Lin, Lingjing Kong, Sebastian U Stich and Martin Jaggi. 'Ensemble distillation for robust model fusion in federated learning'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2351–2363.

[101] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C Lawrence Zitnick. 'Microsoft coco: Common objects in context'. In: *ECCV*. Springer. 2014.

[102] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu and Yi Yang. 'Improving Person Re-identification by Attribute and Identity Learning'. In: *arXiv:1703.07220* (2017).

[103] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao and Yaohai Huang. 'Fair loss: Margin-aware reinforcement learning for deep face recognition'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 10052–10061.

[104] Chunxiao Liu, Shaogang Gong, Chen Change Loy and Xinggang Lin. 'Person re-identification: What features are important?' In: *ECCV*. Springer. 2012.

[105] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan and Xiaogang Wang. 'Hydraplus-net: Attentive deep features for pedestrian analysis'. In: *ICCV*. 2017.

[106] Yang Liu, Qingchao Chen and Andrew Zisserman. 'Amplifying key cues for human-object-interaction detection'. In: *European Conference on Computer Vision*. Springer. 2020, pp. 248–265.

[107] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding and Jungong Han. 'From zero-shot learning to conventional supervised classification: Unseen visual data synthesis'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1627–1636.

[108] Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou and Masashi Sugiyama. 'Unsupervised Federated Learning is Possible'. In: *International Conference on Learning Representations*. 2021.

[109] Chuanchen Luo, Chunfeng Song and Zhaoxiang Zhang. 'Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup'. In: *European Conference on Computer Vision*. Vol. 2. 6. Springer. 2020, p. 7.

[110] Xinhong Ma, Junyu Gao and Changsheng Xu. 'Active Universal Domain Adaptation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 8968–8977.

[111] Rafid Mahmood, Sanja Fidler and Marc T Law. 'Low Budget Active Learning via Wasserstein Distance: An Integer Programming Approach'. In: *arXiv preprint arXiv:2106.02968* (2021).

[112] Xudong Mao, Yun Ma, Zhenguo Yang, Yangbin Chen and Qing Li. 'Virtual mixup training for unsupervised domain adaptation'. In: *arXiv preprint arXiv:1905.04215* (2019).

[113] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang and Houqiang Li. 'CLIP4HOI: Towards Adapting CLIP for Practical Zero-Shot HOI Detection'. In: *Advances in Neural Information Processing Systems* 36 (2024).

[114] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni and Richard Vidal. 'Federated multi-task learning under a mixture of distributions'. In: *Advances in Neural Information Processing Systems* 34 (2021).

[115] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson and Blaise Aguera y Arcas. 'Communication-efficient learning of deep networks from decentralized data'. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[116] Prem Melville and Raymond J Mooney. 'Diverse ensembles for active learning'. In: *ICML*. 2004, p. 74.

[117] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu and Rongrong Ji. 'Active teacher for semi-supervised object detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14482–14491.

[118] Yifei Ming, Hang Yin and Yixuan Li. 'On the impact of spurious correlation for out-of-distribution detection'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9. 2022, pp. 10051–10059.

[119] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa and Hassan Ghasemzadeh. 'Improved knowledge distillation via teacher assistant'. In: *AAAI*. 2020.

[120] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li and Jinwoo Shin. 'Object-aware contrastive learning for debiased scene representation'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12251–12264.

[121] Mazda Moayeri, Phillip Pope, Yogesh Balaji and Soheil Feizi. 'A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19087–19097.

[122] Mehryar Mohri, Gary Sivek and Ananda Theertha Suresh. 'Agnostic federated learning'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4615–4625.

[123] Varun Nair, Javier Fuentes Alonso and Tony Beltramelli. 'Realmix: Towards realistic semi-supervised deep learning algorithms'. In: *arXiv preprint arXiv:1912.08766* (2019).

[124] David Nilsson, Aleksis Pirinen, Erik Gärtner and Cristian Sminchisescu. 'Embodied visual active learning for semantic segmentation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2373–2383.

[125] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma and Yefeng Zheng. 'Multi-anchor active domain adaptation for semantic segmentation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9112–9122.

[126] Sakrapee Paisitkriangkrai, Chunhua Shen and Anton Van Den Hengel. 'Learning to rank in person re-identification with metric ensembles'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1846–1855.

[127] Omkar M Parkhi, Andrea Vedaldi and Andrew Zisserman. 'Deep face recognition'. In: (2015).

[128] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton van den Hengel and Javen Qinfeng Shi. 'Active Learning by Feature Mixing'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12237–12246.

[129] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga and Adam Lerer. 'Automatic differentiation in PyTorch'. In: *NIPS-W*. 2017.

[130] Xingchao Peng, Zijun Huang, Yizhe Zhu and Kate Saenko. 'Federated Adversarial Domain Adaptation'. In: *International Conference on Learning Representations*. 2019.

[131] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen and Song-Chun Zhu. 'Learning human-object interactions by graph parsing neural networks'. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 401–417.

[132] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong and Dacheng Tao. 'Distillation using oracle queries for transformer-based human-object interaction detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19558–19567.

[133] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty and Sethuraman Panchanathan. 'Deep active learning for image classification'. In: *2017 ICIP*. IEEE. 2017.

[134] Daniel A Reid, Mark S Nixon and Sarah V Stevenage. 'Soft biometrics; human identification using comparative descriptions'. In: *IEEE TPAMI* 36.6 (2014).

[135] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi et al. 'Balanced meta-softmax for long-tailed visual recognition'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4175–4186.

[136] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. 'High-resolution image synthesis with latent diffusion models'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[137] Danila Rukhovich and Danil Galeev. 'MixMatch Domain Adaptaion: Prize-winning solution for both tracks of VisDA 2019 challenge'. In: *arXiv preprint arXiv:1910.03903* (2019).

[138] M Saquib Sarfraz, Arne Schumann, Andreas Eberle and Rainer Stiefelhagen. 'A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking'. In: *CVPR*. 2018.

[139] Walter J Scheirer, Neeraj Kumar, Peter N Belhumeur and Terrance E Boult. 'Multi-attribute spaces: Calibration for attribute fusion and similarity search'. In: *CVPR*. IEEE. 2012.

[140] Ozan Sener and Silvio Savarese. 'Active learning for convolutional neural networks: A core-set approach'. In: *arXiv preprint arXiv:1708.00489* (2017).

[141] H Sebastian Seung, Manfred Opper and Haim Sompolinsky. 'Query by committee'. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 287–294.

[142] Claude Elwood Shannon. 'A mathematical theory of communication'. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[143] Yichun Shi and Anil K Jain. 'Boosting unconstrained face recognition with auxiliary unlabeled data'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2795–2804.

[144] Zhiyuan Shi, Timothy M Hospedales and Tao Xiang. 'Transferring a semantic representation for person re-identification and search'. In: *CVPR*. 2015.

[145] Yawar Siddiqui, Julien Valentin and Matthias Nießner. 'ViewAL: Active Learning with Viewpoint Entropy for Semantic Segmentation'. In: *CVPR*. 2020.

[146] Behjat Siddiquie, Rogerio S Feris and Larry S Davis. 'Image ranking and retrieval based on multi-attribute queries'. In: *CVPR*. IEEE. 2011.

[147] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep and Yong Jae Lee. 'Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond'. In: *arXiv preprint arXiv:1811.02545* (2018).

[148] Samarth Sinha, Sayna Ebrahimi and Trevor Darrell. 'Variational adversarial active learning'. In: *ICCV*. 2019.

[149] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis and Wen Gao. 'Multi-task learning with low rank attribute embedding for person re-identification'. In: *ICCV*. 2015.

[150] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao and Qi Tian. 'Deep attributes driven multi-camera person re-identification'. In: *ECCV*. Springer. 2016.

[151] Yifan Sun, Liang Zheng, Weijian Deng and Shengjin Wang. 'Svdnet for pedestrian retrieval'. In: *ICCV*. 2017.

[152] Masato Tamura, Hiroki Ohashi and Tomoaki Yoshinaga. 'Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10410–10419.

[153] Yihe Tang, Weifeng Chen, Yijun Luo and Yuting Zhang. 'Humble teachers teach better students for semi-supervised object detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3132–3141.

[154] Yonglong Tian, Dilip Krishnan and Phillip Isola. 'Contrastive Representation Distillation'. In: *ICLR*. 2019.

[155] Andrew Top, Ghassan Hamarneh and Rafeef Abugharbieh. ‘Active learning for interactive 3D image segmentation’. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011.

[156] Yao-Hung Hubert Tsai, Liang-Kang Huang and Ruslan Salakhutdinov. ‘Learning robust visual-semantic embeddings’. In: *ICCV*. IEEE. 2017.

[157] Eric Tzeng, Judy Hoffman, Kate Saenko and Trevor Darrell. ‘Adversarial discriminative domain adaptation’. In: *CVPR*. 2017.

[158] Mesut Erhan Unal and Adriana Kovashka. ‘Weakly-Supervised HOI Detection from Interaction Labels Only and Language/Vision-Language Priors’. In: *arXiv preprint arXiv:2303.05546* (2023).

[159] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur and Matthew Turk. ‘Attribute-based people search in surveillance environments’. In: *WACV*. IEEE. 2009.

[160] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Yoshua Bengio and David Lopez-Paz. ‘Interpolation consistency training for semi-supervised learning’. In: *arXiv preprint arXiv:1903.03825* (2019).

[161] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz and Yoshua Bengio. ‘Manifold mixup: Better representations by interpolating hidden states’. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6438–6447.

[162] Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars and Xuming He. ‘Weakly-supervised HOI Detection via Prior-guided Bi-level Representation Learning’. In: *arXiv preprint arXiv:2303.01313* (2023).

[163] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic and Heng Tao Shen. ‘Adversarial cross-modal retrieval’. In: *ACM MM*. ACM. 2017.

[164] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang and Lei Zhang. ‘Joint learning of single-image and cross-image representations for person re-identification’. In: *CVPR*. 2016.

[165] Jingya Wang, Xiatian Zhu, Shaogang Gong and Wei Li. ‘Attribute recognition by joint recurrent learning of context and correlation’. In: *ICCV*. 2017.

[166] Jingya Wang, Xiatian Zhu, Shaogang Gong and Wei Li. 'Transferable joint attribute-identity deep learning for unsupervised person re-identification'. In: *CVPR*. 2018.

[167] Kaiye Wang, Ran He, Wei Wang, Liang Wang and Tieniu Tan. 'Learning coupled feature spaces for cross-modal matching'. In: *ICCV*. 2013.

[168] Liwei Wang, Yin Li and Svetlana Lazebnik. 'Learning deep structure-preserving image-text embeddings'. In: *CVPR*. 2016.

[169] Qian Wang and Daniel Kurz. 'Reconstructing Training Data from Diverse ML Models by Ensemble Inversion'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2909–2917.

[170] Shuo Wang, Yuexiang Li, Kai Ma, Ruhui Ma, Haibing Guan and Yefeng Zheng. 'Dual Adversarial Network for Deep Active Learning'. In: 2020.

[171] Weiran Wang, Raman Arora, Karen Livescu and Jeff Bilmes. 'On deep multi-view representation learning'. In: *ICML*. 2015.

[172] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, M Bagheri and R Summers. 'Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases'. In: *IEEE CVPR*. Vol. 7. 2017.

[173] Yu-Xiong Wang, Ross Girshick, Martial Hebert and Bharath Hariharan. 'Low-shot learning from imaginary data'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7278–7286.

[174] Zhiguo Wang, Xintong Wang, Ruoyu Sun and Tsung-Hui Chang. 'Federated Semi-Supervised Learning with Class Distribution Mismatch'. In: *arXiv preprint arXiv:2111.00010* (2021).

[175] Jiaxi Wu, Jiaxin Chen and Di Huang. 'Entropy-based active learning for object detection with progressive diversity constraint'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9397–9406.

[176] Shirley Wu, Mert Yuksekgonul, Linjun Zhang and James Zou. 'Discover and Cure: Concept-aware Mitigation of Spurious Correlation'. In: *arXiv preprint arXiv:2305.00650* (2023).

[177] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu and Cewu Lu. 'Mining cross-person cues for body-part interactiveness learning in hoi detection'. In: *European Conference on Computer Vision*. Springer. 2022, pp. 121–136.

[178] Yongqin Xian, Christoph H Lampert, Bernt Schiele and Zeynep Akata. 'Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly'. In: *IEEE transactions on pattern analysis and machine intelligence* 41.9 (2018), pp. 2251–2265.

[179] Liuyu Xiang and Guiguang Ding. 'Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-tailed Classification'. In: *arXiv preprint arXiv:2001.01536* (2020).

[180] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu and Xinjing Cheng. 'Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 8068–8078.

[181] Fei Yan and Krystian Mikolajczyk. 'Deep correlation for matching images and text'. In: *CVPR*. 2015.

[182] Chenglin Yang, Lingxi Xie, Chi Su and Alan L Yuille. 'Snapshot distillation: Teacher-student optimization in one generation'. In: *CVPR*. 2019.

[183] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu and Ming-Hsuan Yang. 'Federated multi-target domain adaptation'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1424–1433.

[184] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang and Jianhuang Lai. 'Adversarial Attribute-Image Person Re-identification'. In: *IJCAI*. July 2018.

[185] Donggeun Yoo and In So Kweon. 'Learning loss for active learning'. In: *CVPR*. 2019.

[186] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang and Sung Ju Hwang. 'Federated continual learning with weighted inter-client transfer'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12073–12086.

[187] Shan You, Chang Xu, Chao Xu and Dacheng Tao. 'Learning from multiple teacher networks'. In: *23rd ACM SIGKDD*. 2017.

[188] Shan You, Chang Xu, Chao Xu and Dacheng Tao. 'Learning with single-teacher multi-student'. In: *22nd AAAI*. 2018.

[189] Fisher Yu, Vladlen Koltun and Thomas Funkhouser. 'Dilated residual networks'. In: *CVPR*. 2017.

[190] Weiping Yu, Sijie Zhu, Taojiannan Yang and Chen Chen. 'Consistency-based active learning for object detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3951–3960.

[191] Di Yuan, Xiaojun Chang, Qiao Liu, Yi Yang, Dehua Wang, Minglei Shu, Zhenyu He and Guangming Shi. 'Active learning for deep visual tracking'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[192] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha and Qingming Huang. 'State-Relabeling Adversarial Active Learning'. In: *CVPR*. 2020.

[193] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin and David Lopez-Paz. 'mixup: Beyond empirical risk minimization'. In: *arXiv preprint arXiv:1710.09412* (2017).

[194] Li Zhang, Tao Xiang and Shaogang Gong. 'Learning a discriminative null space for person re-identification'. In: *CVPR*. 2016.

[195] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao and Kaisheng Ma. 'Be your own teacher: Improve the performance of convolutional neural networks via self distillation'. In: *ICCV*. 2019.

[196] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo and Xiaoming Jin. 'Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning.' In: *IJCAI*. 2018.

[197] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin and Vikas Chandra. 'Federated learning with non-iid data'. In: *arXiv preprint arXiv:1806.00582* (2018).

[198] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang and Qi Tian. 'Scalable person re-identification: A benchmark'. In: *ICCV*. 2015.

[199] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang and Yi-Dong Shen. 'Dual-Path Convolutional Image-Text Embedding with Instance Loss'. In: *arXiv:1711.05535* (2017).

[200] Zhun Zhong, Liang Zheng, Donglin Cao and Shaozi Li. 'Re-ranking person re-identification with k-reciprocal encoding'. In: *CVPR*. 2017.

[201] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang and Nicu Sebe. 'Openmix: Reviving known knowledge for discovering novel visual categories in an open world'. In: *arXiv preprint arXiv:2004.05551* (2020).

[202] Kaiyang Zhou and Tao Xiang. 'Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch'. In: *arXiv preprint arXiv:1910.10093* (2019).

[203] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro and Tao Xiang. 'Learning Generalisable Omni-Scale Representations for Person Re-Identification'. In: *arXiv preprint arXiv:1910.06827* (2019).

[204] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro and Tao Xiang. 'Omni-Scale Feature Learning for Person Re-Identification'. In: *ICCV*. 2019.

[205] Penghao Zhou and Mingmin Chi. 'Relation parsing neural network for human-object interaction detection'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 843–851.

[206] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A Efros. 'Unpaired image-to-image translation using cycle-consistent adversarial networks'. In: *ICCV*. 2017.

[207] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng and Ahmed Elgammal. 'A generative adversarial approach for zero-shot learning from noisy texts'. In: *CVPR*. 2018.

[208] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei et al. 'End-to-end human object interaction detection with hoi transformer'. In: *CVPR*. 2021, pp. 11825–11834.