Efficient Learning for Fine-grained Recognition

Mingjiang Liang

Doctor of Philosophy

A thesis Submitted for the Degree of Doctor of Philosophy at
University of Technology Sydney in 2025
School of Computer Science, FEIT

Certificate of Original Authorship

I, Mingjiang Liang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Print Name: Mingjiang Liang

Date: 23/06/2025

Acknowledgements

Throughout my doctoral study at the University of Technology Sydney, I have received a great deal of support and assistance.

I would like to express my deepest gratitude to my supervisor, Prof. Wei Liu, for his invaluable guidance, support, and encouragement throughout my doctoral research. His expertise, patience, and insightful feedback have greatly shaped my academic growth and the development of this thesis.

I would also like to thank my co-supervisor, Prof. Jianlong Zhou, for his continuous support, insightful suggestions, and constructive criticism, which have been crucial to the progress of my research.

My heartfelt thanks go to my family for their unwavering love, support, and encouragement during my studies. Their understanding and patience have been a constant source of strength, and I am truly grateful for their belief in me.

Finally, I would like to extend my appreciation to all those who have contributed to my research and journey, whether directly or indirectly.

Publications During Enrolment

Publications related to this thesis:

1. **Mingjiang Liang**, Shaoli Huang, Shirui Pan, Mingming Gong, and Wei Liu. *"Learning Multi-level Weight-centric Features for Few-shot Learning."* Pattern Recognition. (Accepted)

2. **Mingjiang Liang**, Shaoli Huang and Wei Liu. *"Dynamic Semantic Structure Distillation for Low-resolution Fine-grained Recognition."* Pattern Recognition. (Accepted)

3. **Mingjiang Liang**, Shaoli Huang, and Wei Liu. *"A Survey on Fine-Grained Image Classification: Recent Advances, Challenges, and Opportunities."* Submitted to Pattern Recognition. (Under Review)

Publications unrelated to this thesis (* contributed equally):

1. **Mingjiang Liang***, Yongkang Cheng*, Hualin Liang, Shaoli Huang, and Wei Liu. *"RopeTP: Global Human Motion Recovery via Integrating Robust Pose Estimation with Diffusion Trajectory Prior "* IEEE/CVF Winter Conference on Applications of Computer Vision 2025. (Accepted)

2. Jikai Zheng*, **Mingjiang Liang***, Shaoli Huang and Jifeng Ning. *"Exploring the Feature Extraction and Relation Modeling For Light-Weight Transformer Tracking ."* European Conferenceon Computer Vision 2025. (Accepted)

3. Yongkang Cheng*, **Mingjiang Liang***, Shaoli Huang, Gaoge Han, Jifeng Ning and Wei Liu. *"Conditional GAN for Enhancing Diffusion Models in Efficient and Authentic Global Gesture Generation from Audio."* IEEE/CVF Winter Conference on Applications of Computer Vision 2025. (Accepted)

4. Gaoge Han*, **Mingjiang Liang***, Jinglei Tang, Yongkang Cheng, Wei Liu and Shaoli Huang. *"Reindiffuse: Crafting physically plausible motions with reinforced diffusion model."* IEEE/CVF Winter Conference on Applications of Computer Vision 2025. (Accepted)

5. Yongkang Cheng*, **Mingjiang Liang***, Shaoli Huang, Jifeng Ning and Wei Liu. *"Expgest: Expressive speaker generation using diffusion model and hybrid audio-text guidance."* IEEE International Conference on Multimedia and Expo 2025. (Accepted)

Abstract

Fine-grained image classification, which aims to distinguish visually similar subcategories, has gained increasing attention in computer vision. However, compared to standard image classification, fine-grained classification presents unique challenges due to the limited availability of labeled training data and the combination of low inter-class variance with high intra-class variance. This thesis addresses these challenges by exploring efficient learning strategies for deep models in fine-grained recognition.

First, we provide a comprehensive review of existing literature, summarizing key advancements in fine-grained recognition. Through an in-depth analysis of the underlying mechanisms of existing methods, we identify key limitations and draw attention to the critical yet underexplored problem of efficient learning—encompassing both data efficiency and model efficiency.

Second, few-shot learning has emerged as a promising approach to mitigate data inefficiency. However, existing methods often fail to fully leverage the representation power of unseen categories and the weight generation capacity in feature learning, leading to performance bottlenecks. To address this, we propose a multi-level weight-centric feature learning framework. This approach enhances the dual role of the feature extractor in few-shot learning through two key techniques: (1) a weight-centric training strategy, which improves the prototype-ability of features, enabling the construction of more discriminative decision boundaries with only a few samples, and (2) a multi-level feature incorporation mechanism, which integrates mid-level and relation-level information to enhance transferability for novel categories while preserving classification accuracy for base classes. Extensive experiments on low-shot classification benchmarks demonstrate that our method significantly outperforms existing approaches.

Additionally, we tackle the low-resolution classification problem by proposing a dynamic semantic structure distillation framework. Our approach perturbs semantic structures to facilitate knowledge distillation while introducing a decoupled distillation objective to preserve essential part relations. We evaluate our approach on two knowledge distillation tasks: high-to-low resolution and large-to-small model distillation. Experimental results confirm its superiority in low-resolution fine-grained classification and its effectiveness in general image classification.

In summary, this thesis advances fine-grained image classification by introducing novel techniques for few-shot learning and low-resolution knowledge distillation, both of which contribute to improving data and model efficiency. Our findings provide valuable insights into efficient deep-learning strategies for fine-grained recognition.

Contents

List of Figures

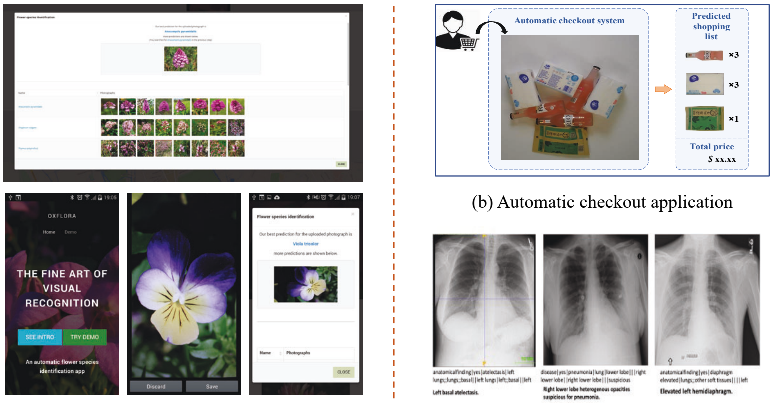
List of Tables

Chapter 1

Introduction

1.1 Background

Image classification has been one of the most vital problems in computer vision. This task aims to determine the object category from images or videos. With the fast progress of machine learning technology, this field has made leaps and bounds in recent years [4–8]. For example, deep learning-based approaches have enabled the machine to exceed human-level performance on large-scale image datasets [8].


OXFLORA
THE FINE ART OF VISUAL RECOGNITION
SEE INTRO
TRY DEMO
An automatic flower species identification app
Discard
Save
Flower species identification
Automatic checkout system
Predicted shopping list
×3
×3
×1
Total price
$ xx.xx
Left basal atelectasis.
Right lower lobe heterogenous opacities suspicious for pneumonia.
Elevated left hemidiaphragm.

(b) Automatic checkout application

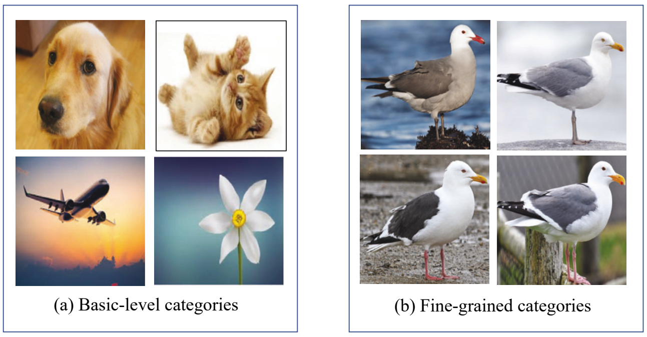
(a) The flower recognition system

(c) Chest X-Ray recognition

Figure 1.1: Some example applications for fine-grained recognition. (a) A system for flower species classification developed by Chai [1]; (b) An automatic checkout application that can recognize each product and show a complete shopping list with the total price introduced by Wei et al. [2]; (c) An IBM AI system that can understand Chest X-Ray images at expert levels [3].

As an emerging subfield of image classification, fine-grained object recognition has gained increasing attention in the last few years [9–16] due to its wide variety of application prospects. Examples of applications such as automatic species recognition systems [1], diagnostic medical imaging [2], and retail product recognition [3] are shown in Figure 1.1.

Compared with object recognition problems aiming to classify ordinary category objects such as cats, dogs, and cars, fine-grained vision categorization (FGVC) concerns distinguishing subordinate categories like bird species or car models. As illustrated in Figure 1.2, ordinary categories usually have noticeably discriminative appearances. However, there are subtle differences between fine-grained classes that are difficult to distinguish by laypeople.


(a) Basic-level categories

(b) Fine-grained categories

FIGURE 1.2: An illustration of basic-level categories and fine-grained categories

These data characteristics make FGVC more challenging than standard object classification. First, since data labeling requires professional domain knowledge, the dataset used for model learning tends to be small and costly. Moreover, the fine-grained datasets exhibit high intra-class but low inter-class variance. As Figure 1.3 shows, the first row of images from the same category look different because of various viewpoints, poses and colours. On the other hand, images in the first column have similar appearances yet belong to different species and are hard to distinguish by non-expert people.

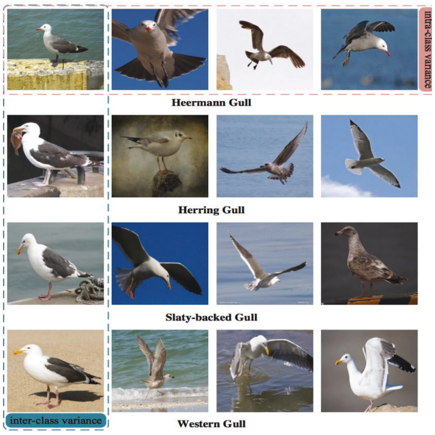


FIGURE 1.3: An illustration of the main challenge in FGVC

Numerous works [9–19] have been devoted to these challenges in the past two decades. Early research employed a combination of computer and human feedback to classify the fine-grained categories. Due to the imperfect computer vision algorithms, it is difficult to capture reliable discriminative features, so embedding human knowledge into computer vision algorithms is effective. These works focus on designing the human-computer interaction process to improve the system's performance and reduce human efforts. Although these works have proved their effectiveness, human feedback is an indispensable part of the algorithm, leading to low system automation.

With the rise of the deep learning technique, the increasingly powerful capabilities offered by deep convolutional neural networks (CNN) [4, 5, 7, 8] have fostered more advanced and automatic algorithms [11, 14–16, 18] in fine-grained recognition. This line of literature mainly employs learnable CNN features in the classification framework instead of traditional hand-crafted features such as SIFT [20], HOG [21] or Fisher Vector [22]. One of the most significant advantages of utilizing CNN is that it can automatically and adaptively learn hierarchical features from images. In contrast, traditional feature descriptors require expert-level knowledge in design and usually fail to extract complex information from images. In addition, CNN-based algorithms have been widely demonstrated to show exceptional superior performance over conventional approaches [11, 14–16, 18].

1.2 Research Problems and Objectives

As mentioned above, fine-grained recognition is different from general object recognition, as the discriminative differences of categories are subtle and hard to distinguish. Although remarkable success has been achieved in fine-grained classification, there are still some new problems in this field. They are reflected in the following aspects.

Challenges in Data Efficiency. Fine-grained visual classification (FGVC) tasks are often hindered by the dual challenges of insufficient labeled data and computational constraints. These challenges manifest in two major aspects.

Firstly, existing networks typically need massive parameters to increase the model's capacity and computing power and enable the model to complete more complex learning tasks. Thus, large-size, well-labeled datasets are critical for fine-grained recognition. Such as CUB200-2010 [23] is one of the most influential benchmark datasets, containing 200 bird categories with 15 training images and 15+ testing images for each class. In 2011, the CUB200-2011 dataset [24] doubled the number of training and testing images to 30 per category. After that, the author again enlarged the amount to form a new high-quality data set called NABirds [25], which included 555 bird categories and more than 100 images for each class. The increasing size of the dataset promotes the development of the fine-grained recognition field based on the deep learning method and facilitates the comparison of the performance of different approaches [26]. However, if we want to classify some rare species, it is difficult to obtain the data. In this data-sparse situation, these methods usually fail to work well.

Secondly, high-resolution as model input has become a trend for optimizing classification performance, as high-resolution images contain richer features and subtle differences [27–29]. The earlier work, such as DPD + DecAF [30], Deep LAC [31], Multi-proposal [32] and Part R-CNN [33] mainly used the resolution of 224 × 224. Since 2016, most frameworks [14, 15, 34]have adopted the resolution of 448 × 448 to achieve higher precision. However, taking advantage of high resolution will lead to large GPU memory consumption and high computation cost [35]. Meanwhile, the application of the model will be limited. Because in our real world, low-resolution images are often more common, and these methods do not perform as effectively in such scenarios. In contrast, low-resolution images are accessible to missing important details, bringing great challenges to fine-grained recognition, and still are unexplored.

In response to the weakness of existing works discussed above, this thesis addresses the following research questions:

- What are the main methods in the field of fine-grained classification? What are their advantages and limitations? What challenges does this field face and how is it likely to evolve in the future?
- Given a very small amount of training data, how do we make the best use of these limited data to train a model with good performance, even comparable to a model trained on sufficient data?
- How to learn to extract discriminative part features from low-resolution images? Then improve the classification performance on low-resolution images.

By considering the above questions, my thesis will achieve the following objectives.

- Develop a Multi-level Weight-centric feature learning method to improve the representation power for a few amounts and unseen samples and solve the problem of fine-grained recognition in few-shot learning.

- Design a model that maintains high performance under low-resolution image input for fine-grained recognition through knowledge distillation.

In support of these objectives, this thesis also presents a comprehensive review of existing literature, which not only categorizes and summarizes existing methods based on their technical approaches, but also outlines the current challenges in this field and provides insights into future research directions.

1.3 Research Significance

It is well known that fine-grained recognition has a wide range of applications, including Smart Retail, recognizing retail products in supermarkets; Medical diagnoses, classifying X-ray images; Ecological Protection, and identifying different species of organisms. The future impact and significance of this thesis will be analyzed from the following two aspects:

Firstly, a deep neural network as a data-hungry model basically needs massive parameters and often fails to work well in data-scarce environments, which challenge us in the real world. This thesis provides a few-shot learning framework to classify a few images of new categories and then applied to the fine-grained recognition task and achieve outstanding performance. With this method, we effectively address the problem of limited training datasets and simultaneously enhance the generalizability of the model. Compared with those approaches of expanding new benchmarks, we save plenty of labor resources and develop a new perspective for the combination of fine-grained recognition and few-shot learning.

Secondly, fine-grained recognition is a research topic closely related to practical applications, aiming to resolve factual issues. However, the benchmark datasets used in current studies generally have the characteristics of prominent foreground objects and single backgrounds. In fact, it is not common in the real world. Hence, we need to consider lighting, blur, occlusion, low-resolution, etc., factors that are lacking in most existing systems to increase the application of the model in natural scenes. From the perspective of model input and complexity, our research will combine some novel areas of knowledge and develop a lightweight classification system for low-resolution images, which has a further impact.

1.4 Thesis Outline

The remainder of the outline of this thesis is as follows:

Chapter 2: This chapter reviews fine-grained recognition approaches, including part alignment, high-order feature learning, and transformers. It also examines dataset evaluation, covering biases, ethical concerns, and performance evaluation. The discussion provides a foundation for understanding current methods and identifying research gaps.

Chapter 3: This chapter proposes a multi-level weight-centric feature learning to give full play to feature extractor's dual roles in few-shot learning, and evaluates the approach to low-shot classification benchmarks.

Chapter 4: This chapter presents a dynamic semantic structure distillation learning framework and reports the experimental validation results.

Chapter 5: This chapter discusses the strengths, limitations, and broader implications of the research. Highlights key contributions, including a comprehensive survey, advances in few-shot learning, and improvements in low-resolution fine-grained recognition, while also addressing existing challenges and outlining future research directions.

Chapter 6: This chapter concludes the thesis and sheds light on future research work on fine-grained recognition.

Chapter 2

Literature Review

2.1 Fine-grained Recognition Approaches

We present an exhaustive exploration of the methods in fine-grained recognition, and organize the literature into seven representative paradigms: 1) Human in the Loop; 2) Part Alignment-Based; 3) High-Order Feature Learning; 4) Multiple Granularity Features; 5) Data Augmentation; 6) Rich Representation Learning; and 7) Transformer-Based Approaches. Rather than claiming a strict taxonomy from an algorithmic perspective, this categorization reflects major research trends and methodological shifts in the field from 2010 to 2023, as visualized in Fig. 2.1. Some paradigms may partially overlap in the underlying techniques, for example, Transformer-based models often incorporate enhanced representation learning strategies, but are separated here based on their primary methodological focus or historical emergence.This structured organization allows us to trace how the field has evolved and to highlight the innovations that have driven performance improvements in FGVC tasks. A summary comparison of the reviewed works is provided in Table 2.1.

We carried out a comprehensive search for relevant articles across multiple databases, including IEEE Xplore, ACM Digital Library, and computer vision conference proceedings. Using search terms such as ”fine-grained recognition”, ”fine-grained visual categorization”, and ”subcategory object recognition”, we obtained more than 500 articles. Subsequently, we screened these articles for relevance, concentrating on those that dealt with fine-grained recognition methods and made contributions to the computer vision field. After this screening, around 200 articles remained. Due to space limitations, we further refined our selection based on the Google citation count (preferring higher counts) along with the diversity and representativeness of the method types. Eventually, we retained approximately 90 articles for in-depth analysis.

The field of fine-grained categorization (FGVC) emerged in 2006. At that time, Nilsback and Zisserman [36] introduced a dataset comprising 17 flower species. This dataset posed a challenge to earlier classification methods because of its high intra-class variance and low inter-class variance. Nilsback's system incorporated Bag-of- Words [36] together with handcrafted features like SIFT [20] and HOG [21] to outline flower contours and colors, emphasizing the importance of segmentation before classification. Later, in 2009, [17] applied flower alignment to pre-segmented flowers, which is relevant to the subsequent paradigms we will discuss.

The first paradigm we consider is the **Human In The Loop** paradigm. As depicted in Fig. 2.2, this approach combines human feedback with machine learning systems to enhance decision-making in FGVC. The main focus of the methods within this paradigm is to integrate human input, such as user-indicated part locations or similarity comparisons. This integration is crucial to improving the feature extraction and classification processes, as demonstrated in the works of [9, 37–40].

Next, we have the **Part Alignment-based** paradigm. This paradigm operates without relying on human involvement. Focuses on aligning images or discriminative features based on shared parts within the same basic-level category. There are two main types of techniques within this paradigm: strongly supervised part alignment, as seen in the work of [10, 11, 31, 33, 41, 42], and weakly supervised part alignment, as in [43–47].

Moving on, the **High-order Feature Learning** paradigm aims to refine the representation of the image. It employs advanced techniques such as bilinear pooling and multi-granularity feature representation. These techniques are used to capture intricate second-order statistical information and maintain the structural integrity of feature maps, as shown in the research of [13, 48–50].

Another important paradigm is the **Multiple Granularity Features** paradigm that focuses on capturing discriminative patterns at various levels of detail to improve classification performance. It includes several approaches such as semantic guided methods, input-driven methods, and expert models for different granularity levels, as described in [14–16, 51].

The **Data Augmentation** paradigm is centered around increasing the quantity and diversity of data. This is achieved through various techniques such as iterative annotation, image blending, and part swapping. By doing so, the generalization capabilities of the model can be improved, as evidenced by the works of [52–54].

The **Enhanced Representation Learning** paradigm involves approaches that explore advanced techniques. These techniques not only take advantage of expanded training data and its diversity but also introduce increased task complexity to enhance the

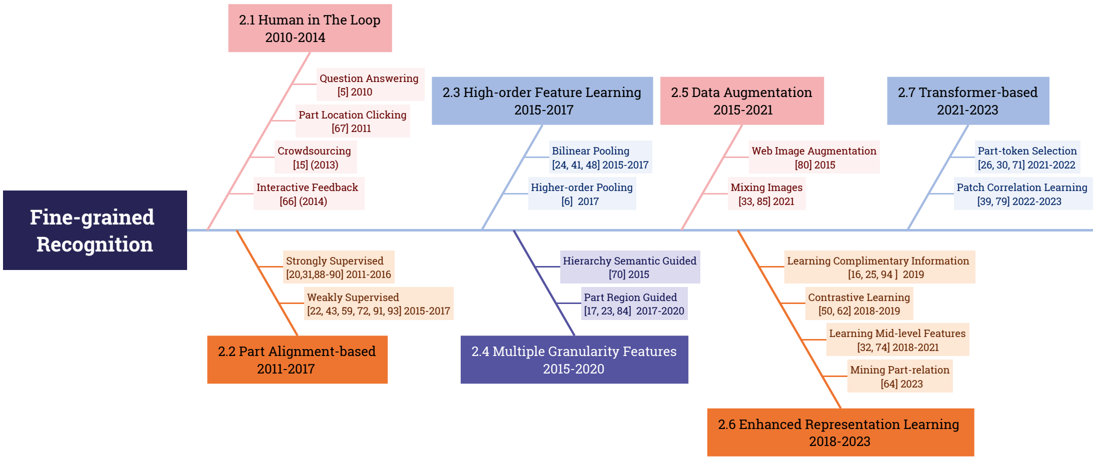


FIGURE 2.1: Overview of the evolution and categorization of techniques in fine-grained image recognition from 2010 to 2023.

model's learning capabilities. This includes aspects such as harnessing complementary information, contrastive learning, incorporating mid-level features, and probing part-relation, as shown in [18, 19, 34, 55–58].

Finally, we have the **Transformer-based Approaches** paradigm. This paradigm utilizes transformer-based techniques to identify complex relationships within images and enhance classification accuracy. The methods within this paradigm can be further divided into token selection in part, as seen in [59–61], and correlation learning in part, as in [62, 63].

Based on the reviewed literature, we identify major challenges and research opportunities in FGVC, while also discussing the limitations of existing approaches. This review outlines the current landscape of the field and highlights potential directions for future research.

Method Category	Methodological Features	Advantages
Human In The Loop	Combines human feedback with machine learning. Uses human input for feature extraction	Improves decision-making, addresses dataset issues, reduces human effort in some cases
Part Alignment-based (Strongly Supervised)	Relies on annotated datasets for detailed part localization models	Accurate part localization, enhanced recognition accuracy
Part Alignment - based (Weakly Supervised)	Automatically discovers parts using weak supervision without detailed annotations	Less labor-intensive, useful for large datasets, can challenge strongly-supervised methods
High - order Feature Learning	Bilinear pooling and polynomial kernel-based methods for feature representation	Refines image representation, reduces the need for strong annotations, captures multi-order statistics
Multiple Granularity Features	Various methods to extract features at different granularities	Enhances classification performance by capturing diverse details
Data Augmentation	Iterative annotation, part swapping and special blending methods	Increase data quantity and diversity, address data limitations and noise issues
Enhanced Representation Learning - Harnessing Complementary Information	Extracts and integrates complementary features using different networks	Improves feature robustness and representation
Enhanced Representation Learning - Contrastive Learning	Uses relationships between features/objects for learning	Refines feature representation, mitigate redundancy
Enhanced Representation Learning - Embracing Mid-level Features	Refines mid-level representation addresses noise sensitivity	Accentuates subtle distinctions, improves model robustness
Enhanced Representation Learning - Probing Part - Relation	Analyzes relationships between object parts for classification	Improves classification in detailed tasks
Transformer-based Approaches - Part Token Selection	Selects and emphasizes relevant image parts using ViT and its enhancements	Enhances classification accuracy and computational efficiency
Transformer-based Approaches - Part Correlation Learning	Learns and represents relationships between object parts	Improves holistic image understanding and classification accuracy

Table 2.1: Comparison of Reviewed Works

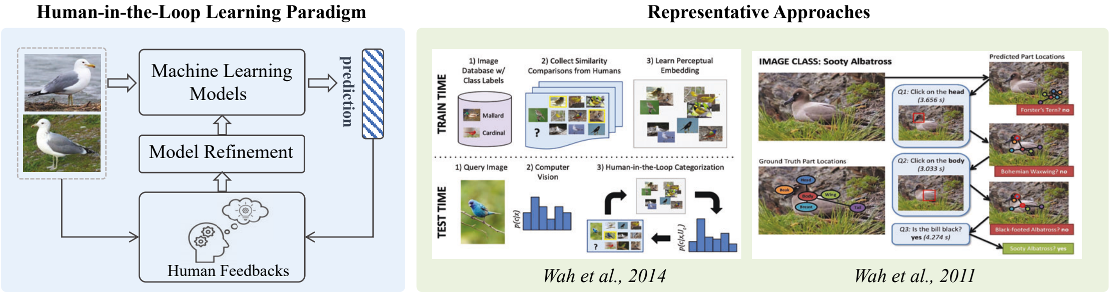

Human-in-the-Loop Learning Paradigm
Machine Learning Models
prediction
Model Refinement
Human Feedbacks
Representative Approaches
TRAIN TIME
1) Image Database w/ Class Labels
Mallard
Cardinal
2) Collect Similarity Comparisons from Humans
?
3) Learn Perceptual Embedding
TEST TIME
1) Query Image
2) Computer Vision
$p(c|x)$
3) Human-in-the-Loop Categorization
$p(c|x,U_t)$
Wah et al., 2014
IMAGE CLASS: Sooty Albatross
Predicted Part Locations
Q1: Click on the head (3.656 s)
Forster's Tern? no
Ground Truth Part Locations
Head
Beak
Body
Wing
Breast
Tail
Q2: Click on the body (3.033 s)
Bohemian Waxwing? no
Q3: Is the bill black? yes (4.274 s)
Black-footed Albatross? no
Sooty Albatross? yes
Wah et al., 2011

Figure 2.2: The human-in-the-loop learning paradigm and the representative approaches.

2.1.1 Human In The Loop

Fine-grained visual categorization is a challenging task for both humans without expertise and machines. The "human-in-the-loop" approach in artificial intelligence research combines human feedback and machine learning systems to improve decision-making. Branson et al. [9] proposed an interactive hybrid human-computer framework that leverages both human and machine strengths in image object recognition. This framework enhances performance using user input while reducing human effort through computer vision. Distinct from the Botanist's Field Guide [64], it is designed for laypeople and accepts unrestricted system input. The framework queries information at runtime, approximating the category of an input image and prompting users with questions. The iterative process of maximizing Information Gain, refining assumptions, and presenting subsequent questions continues until classification is achieved. This human-in-the-loop learning paradigm and representative approaches are visually summarized in Fig. 2.2.

In contrast to Branson's work [9], Wah et al. [37] integrated human input, specifically user-indicated part locations, with part-based and attribute-based computer vision techniques. This augmented object part localization and image classification. Specifically, the algorithms initially estimate the probable location of parts in the image. Users then specify these locations (e.g., head or body), which, in turn, modifies the probability distribution for other part locations. This process persists until image classification is finalized.

Given the scarcity and high cost of fine-grained datasets, Deng et al. [38] devised an online game, "bubble," to gather large-scale data affordably. This game aids computers in identifying discriminating features for pinpointing fine-grained categories. This innovative crowdsourcing method operates twofold: collecting class labels and additional data through the game and integrating with detailed hints to boost performance.

While traditional Fine-Grained Visual Categorization (FGVC) systems depended on expertly crafted attribute and part vocabularies, certain categories, such as chairs or paintings, defy comprehensive vocabulary creation. Addressing this, Wah et al. [39] presented a human-interactive system grounded on perceptual similarity rather than attribute vocabularies. This system cultivates a perceptual embedding from human similarity comparisons during training. At test time, it amalgamates the learned embedding with user feedback and a computer vision algorithm to determine image classes.

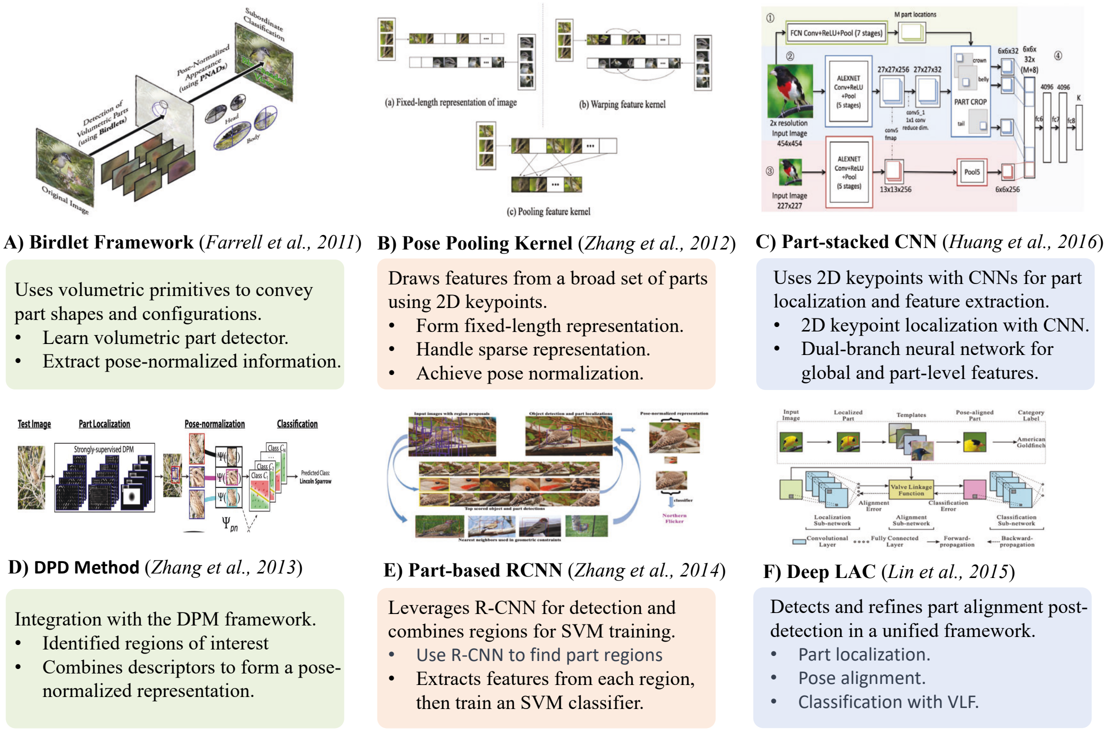
Feature extraction remains pivotal in fine-grained recognition. Early studies [9, 37, 38] harnessed human input to bolster computer vision algorithms in distinguishing features, thus enhancing performance. With the emergence of Convolutional Neural Networks (CNN), many researchers have capitalized on CNN features. However, deep model training demands vast datasets. To address this, Cui et al. [40] introduced a bootstrapping-based framework to generate expansive training datasets. This model couples a pre-trained metric learning model with human responses, iteratively producing high-confidence labeled samples for dataset expansion. Notably, this metric learning model, grounded on a deep CNN, is trained using triplet loss. A drawback, however, is its time-consuming training process.

2.1.2 Part Alignment-based

Although the human-in-the-loop paradigm provides an effective solution for fine-grained recognition, it inherently requires human feedback, which compromises the level of automation in the recognition system. Moreover, the reliance on human interaction escalates with the expansion of the dataset size, rendering these methods less practical for large-scale applications due to the increased demand for human effort.

Consequently, subsequent research has pivoted towards developing fine-grained classification approaches that operate independently of human involvement. An essential branch in this domain is the method based on part alignment, which has garnered significant interest. Part alignment involves arranging images or discriminative features under shared parts within the same basic-level category, facilitating a detailed comparison critical to distinguishing between closely related categories. In addition, this approach mitigates the variances induced by pose, background, and viewing angles that can obscure similarities between images of the same class.

Currently, part alignment techniques can be classified into two main streams: strongly-supervised [10, 11, 33, 41, 42] and weakly-supervised based part localization [43–46, 65, 66]. Strongly supervised approaches operate under the assumption that datasets come with annotated parts, from which a part localization model can be directly learned.



A) Birdlet Framework (*Farrell et al., 2011*)

Uses volumetric primitives to convey part shapes and configurations.
- Learn volumetric part detector.
- Extract pose-normalized information.

B) Pose Pooling Kernel (*Zhang et al., 2012*)

Draws features from a broad set of parts using 2D keypoints.
- Form fixed-length representation.
- Handle sparse representation.
- Achieve pose normalization.

C) Part-stacked CNN (*Huang et al., 2016*)

Uses 2D keypoints with CNNs for part localization and feature extraction.
- 2D keypoint localization with CNN.
- Dual-branch neural network for global and part-level features.

D) DPD Method (*Zhang et al., 2013*)

Integration with the DPM framework.
- Identified regions of interest
- Combines descriptors to form a pose-normalized representation.

E) Part-based RCNN (*Zhang et al., 2014*)

Leverages R-CNN for detection and combines regions for SVM training.
- Use R-CNN to find part regions
- Extracts features from each region, then train an SVM classifier.

F) Deep LAC (*Lin et al., 2015*)

Detects and refines part alignment post-detection in a unified framework.
- Part localization.
- Pose alignment.
- Classification with VLF.

Figure 2.3: Typical methods for Strongly supervised Part Alignment.

However, weakly supervised methods rely solely on class labels to identify discriminative parts without explicit part annotations. In the following parts, a detailed exploration of these two methodologies will be provided.

Strongly Supervised Part Alignment. Following the exploration of part alignment methods in fine-grained visual categorization, we turn our attention to strongly supervised techniques. These methods rely on precise annotations within datasets to learn part localization models, often with great detail and specificity. A visual representation of typical methods used in strongly supervised part alignment is provided in Fig. 2.3.

In 2011, Farrell et al. [10] introduced a novel pose-normalized feature extraction method, based on Poselets [67]. This method, coined as "birdlet," employs volumetric primitives as templates to detail part shapes and configurations. The process begins with the learning of a volumetric part detector to model categories using geometric primitives, followed by the extraction of pose-normalized appearance and shape information to enhance fine-grained recognition accuracy.

However, the "birdlet" framework focused on head and body features, overlooking other discriminative areas like tails or wings. To expand on this, Zhang et al. [41] developed the Pose-Pooling Kernel, extracting features from a broader array of parts. By utilizing 2D keypoints, this method offers a more dimensionally economical feature representation, is streamlined for pose normalization, and requires less annotation labor than 3D methods.

In addition to contributing to this line of research, Huang et al. [11] employed 2D key points in conjunction with deep convolutional neural networks (CNNs) to localize and extract characteristics. Their two-stage approach begins with keypoint localization using a fully convolutional network, followed by feature segmentation and integration using a dual-branch CNN.

Other researchers, such as Zhang et al. [42] and Zhang et al. [33], have utilized bounding boxes for robust part detection, employing detectors to identify part regions and collect features. The DPD method by Zhang et al. [42] and the part-based RCNN by Zhang et al. [33] are prime examples of this approach, leveraging established object detectors to obtain a pose-normalized representation for classification.

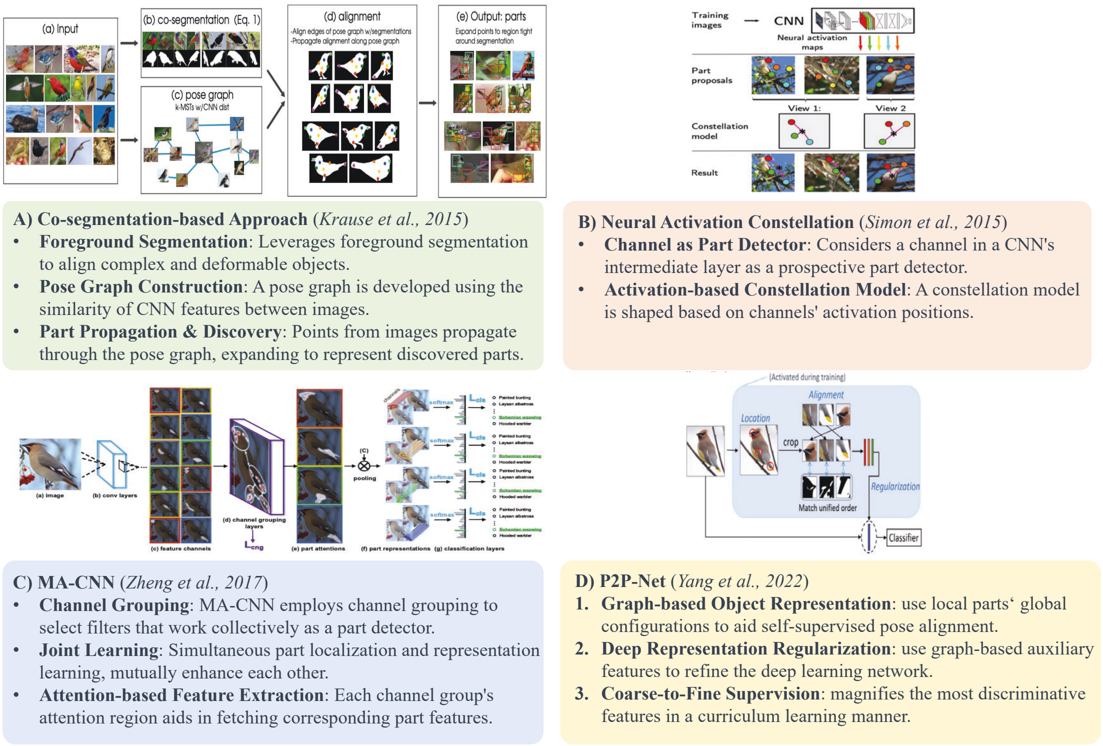
Taking a distinct route, Lin et al. [31] introduced Deep LAC, a method that refines the alignment of the parts after detection and integrates location, pose alignment, and classification. The unique Valve Linkage Function within Deep LAC dynamically adjusts part localization to minimize classification and alignment errors.

Weakly Supervised Part Alignment. Although strongly supervised methods are proficient in predicting accurate part localization, they are contingent on detailed part-level annotations, such as bounding boxes or key points. However, acquiring such annotations is notably labor-intensive, particularly for large datasets. Thus, a shift in research focus has been observed towards approaches that can automatically discover parts using weak supervision, which aims to localize parts without detailed annotations.

Typical weakly supervised part alignment methods are illustrated in Fig. 2.4. Among these methods is the one by Krause et al. [43], which employs a co-segmentation approach to achieve part alignment. By creating a pose graph and propagating key points across images, this method delineates part regions with a degree of accuracy that challenges its strongly supervised counterparts.

Innovative strategies like the Neural Activation Constellation by Simon and Rodner [44] exploit the activations within CNN layers to mine convolutional filters that serve as part detectors. Similarly, the framework introduced by Zhang et al. [45] uses deep filter responses to identify parts, further encoding them for classification purposes.

Yet, the fragmentation of parts detected through individual channel activations is a known limitation. To overcome this, the MA-CNN model proposed by Zheng et al. [46] utilizes channel grouping for collaborative part detection, enhancing the robustness of the parts identified. This method underscores the synergy between part localization and feature learning, demonstrating improved performance through this dual focus.



A) Co-segmentation-based Approach (*Krause et al., 2015*)

- **Foreground Segmentation**: Leverages foreground segmentation to align complex and deformable objects.
- **Pose Graph Construction**: A pose graph is developed using the similarity of CNN features between images.
- **Part Propagation & Discovery**: Points from images propagate through the pose graph, expanding to represent discovered parts.

B) Neural Activation Constellation (*Simon et al., 2015*)

- **Channel as Part Detector**: Considers a channel in a CNN's intermediate layer as a prospective part detector.
- **Activation-based Constellation Model**: A constellation model is shaped based on channels' activation positions.

C) MA-CNN (*Zheng et al., 2017*)

- **Channel Grouping**: MA-CNN employs channel grouping to select filters that work collectively as a part detector.
- **Joint Learning**: Simultaneous part localization and representation learning, mutually enhance each other.
- **Attention-based Feature Extraction**: Each channel group's attention region aids in fetching corresponding part features.

D) P2P-Net (*Yang et al., 2022*)

1. **Graph-based Object Representation**: use local parts' global configurations to aid self-supervised pose alignment.
2. **Deep Representation Regularization**: use graph-based auxiliary features to refine the deep learning network.
3. **Coarse-to-Fine Supervision**: magnifies the most discriminative features in a curriculum learning manner.

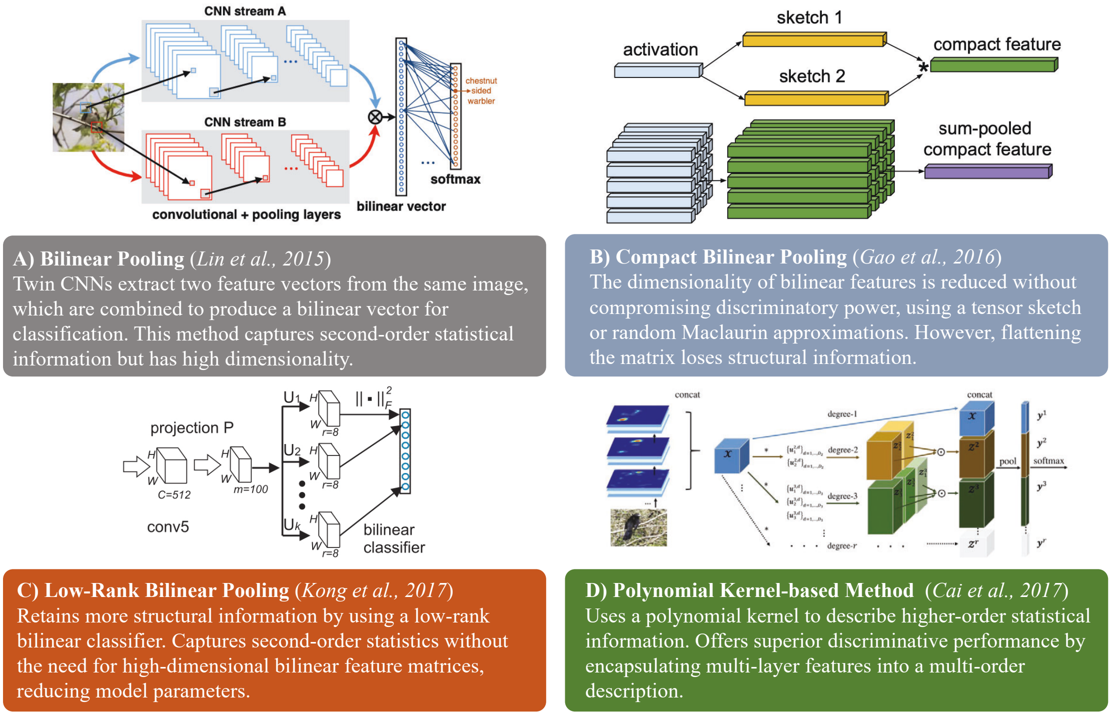
Figure 2.4: Typical Weakly Supervised Part Alignment Approaches.

At the frontier of these advancements is P2P-Net [47], a method that capitalizes on graph-centric object representations to infer local parts' global configurations. This self-supervised approach not only aids in pose alignment but also serves as a feature regularizer within deep learning networks, thereby augmenting the accuracy of fine-grained object classification.

2.1.3 High-order Feature Learning

As we pivot from foundational feature extraction techniques towards more nuanced methods, high-order feature representation emerges as a potent approach to refine image representation, alleviating the need for strong annotations. The typical approaches are illustrated in Fig. 2.5. Among these advanced techniques, bilinear pooling has proven to be particularly effective for fine-grained recognition. It involves the utilization of dual CNNs to extract feature vectors from an image. These vectors are then merged into a bilinear vector, capturing intricate second-order statistical information, as elucidated by Lin et al. [48]. The resultant high-dimensional space derived from this process enriches local features, thus enhancing classification performance.

Nevertheless, bilinear pooling is not without limitations, primarily due to the high dimensionality of the resulting features. This results in a significantly increased parameter



A) Bilinear Pooling (*Lin et al., 2015*)
Twin CNNs extract two feature vectors from the same image, which are combined to produce a bilinear vector for classification. This method captures second-order statistical information but has high dimensionality.

B) Compact Bilinear Pooling (*Gao et al., 2016*)
The dimensionality of bilinear features is reduced without compromising discriminatory power, using a tensor sketch or random Maclaurin approximations. However, flattening the matrix loses structural information.

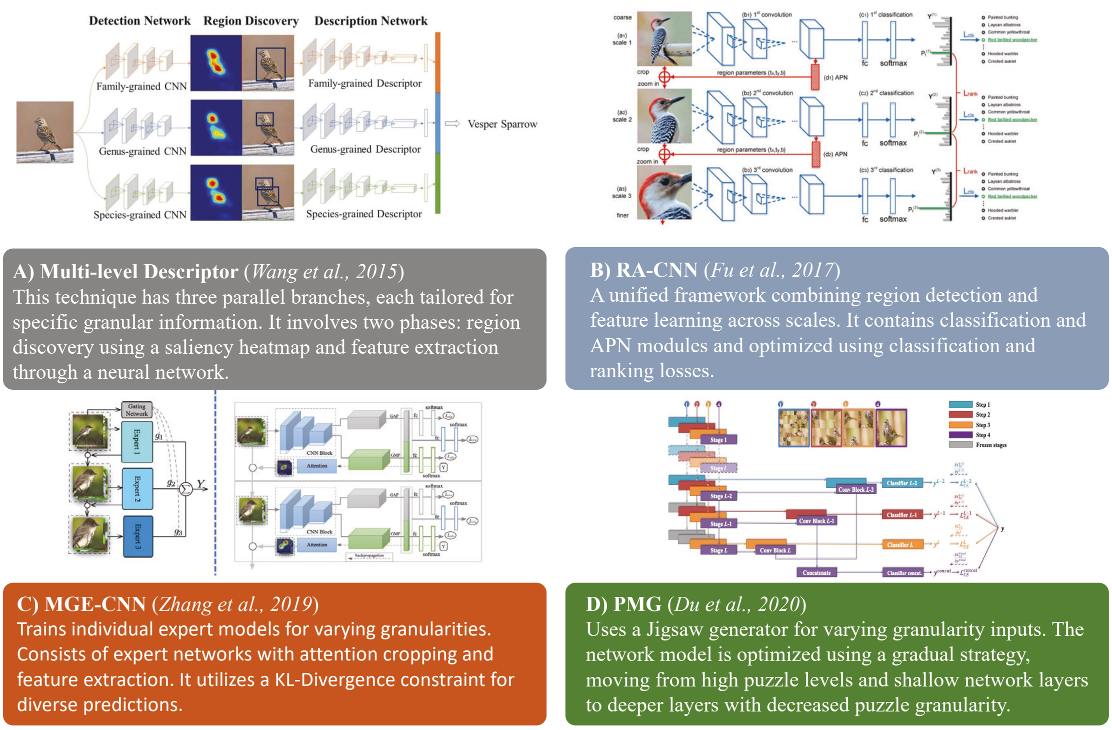
C) Low-Rank Bilinear Pooling (*Kong et al., 2017*)
Retains more structural information by using a low-rank bilinear classifier. Captures second-order statistics without the need for high-dimensional bilinear feature matrices, reducing model parameters.

D) Polynomial Kernel-based Method (*Cai et al., 2017*)
Uses a polynomial kernel to describe higher-order statistical information. Offers superior discriminative performance by encapsulating multi-layer features into a multi-order description.

Figure 2.5: Representative High-order Feature Extraction Techniques.

count for the classifier, which presents computational challenges. In response to this, Gao et al. [49] developed a compact bilinear pooling technique, which leverages tensor sketch and random Maclaurin methods to compress the dimensions of the features while preserving their discriminative strength. However, both the original and compact bilinear pooling methods have the drawback of losing structural information due to vectorization of the pooled matrix [48, 49]. A novel workaround presented by Kong and Fowlkes [13] is the low-rank bilinear pooling method, which employs a low-rank classifier to retain second-order statistics without the computational overhead of high-dimensional matrices.

Expanding the scope beyond bilinear techniques, Cai et al. [50] introduced a high-order statistical representation based on the polynomial kernel. This innovative method concatenates feature maps from different CNN layers and then subjects them to a polynomial transformation. The transformed features, processed through various degrees of a polynomial kernel, are aggregated into a final feature vector for classification. This approach not only preserves the structural integrity of the feature maps but also capitalizes on a multilayer, multiorder statistical information, culminating in a representation with enhanced discriminative capabilities.



A) Multi-level Descriptor (*Wang et al., 2015*)
This technique has three parallel branches, each tailored for specific granular information. It involves two phases: region discovery using a saliency heatmap and feature extraction through a neural network.

B) RA-CNN (*Fu et al., 2017*)
A unified framework combining region detection and feature learning across scales. It contains classification and APN modules and optimized using classification and ranking losses.

C) MGE-CNN (*Zhang et al., 2019*)
Trains individual expert models for varying granularities. Consists of expert networks with attention cropping and feature extraction. It utilizes a KL-Divergence constraint for diverse predictions.

D) PMG (*Du et al., 2020*)
Uses a Jigsaw generator for varying granularity inputs. The network model is optimized using a gradual strategy, moving from high puzzle levels and shallow network layers to deeper layers with decreased puzzle granularity.

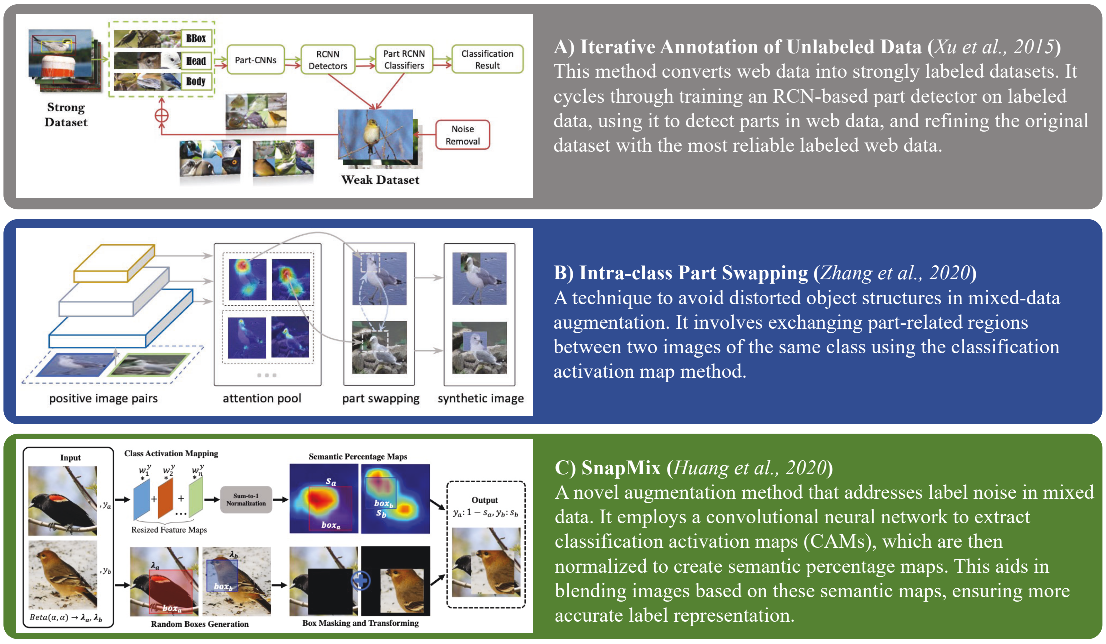
Figure 2.6: Multi-Granularity Feature Representation Methods.

2.1.4 Multiple Granularity Features

Advancing our exploration into sophisticated feature representation techniques, we dive into the realm of multi-granularity feature representation, a proven approach to increase the efficacy of fine-grained recognition tasks [14–16, 51, 68]. Capturing discriminative patterns across various levels of detail, this approach is pivotal in enhancing classification performance. The diverse methodologies employed in multi-granularity feature representation are depicted in Fig. 2.6

In the sphere of semantic-guided methods, hierarchical labels are employed to extract information at different granularities. A notable example is the Multiple Granularity Descriptors by Wang et al. [51], which utilize additional label hierarchies to enrich the representation of features. This method consists of parallel branches, each designed to capture granular information at specific levels, integrating region discovery with feature extraction.

In contrast, input-driven methods, such as the RA-CNN (Recurrent Attention Convolutional Neural Network) proposed by Fu et al. [14], create separate granular inputs to derive features at multiple granularities. This framework intertwines region detection and feature learning, optimizing the model with both classification and ranking losses to ensure a hierarchy in the granularity of predictions.



A) Iterative Annotation of Unlabeled Data (*Xu et al., 2015*) This method converts web data into strongly labeled datasets. It cycles through training an RCN-based part detector on labeled data, using it to detect parts in web data, and refining the original dataset with the most reliable labeled web data.

B) Intra-class Part Swapping (*Zhang et al., 2020*) A technique to avoid distorted object structures in mixed-data augmentation. It involves exchanging part-related regions between two images of the same class using the classification activation map method.

C) SnapMix (*Huang et al., 2020*) A novel augmentation method that addresses label noise in mixed data. It employs a convolutional neural network to extract classification activation maps (CAMs), which are then normalized to create semantic percentage maps. This aids in blending images based on these semantic maps, ensuring more accurate label representation.

Figure 2.7: Data Augmentation Techniques in Fine-Grained Data.

Another distinctive approach is MGE-CNN (Mixture of Granularity-Specific Experts) by Zhang et al. [15], which trains individual expert models for different levels of granularity. Each expert in this network is designed to focus on specific regions, thereby enabling finer-grained data processing, and the outputs are integrated using a gating network.

Additionally, the Progressive-Multi-Granularity (PMG) method introduced by Du et al. [16] represents an innovative implementation of Jigsaw Patches. This approach employs a Jigsaw generator to create images with varying puzzle levels, symbolizing inputs at different granularities. The network is trained progressively, starting with shallow layers for higher puzzle levels and gradually involving deeper layers as the granularity decreases, thereby effectively combining features from multiple scales.

2.1.5 Data Augmentation

In the context of fine-grained recognition, data augmentation plays a crucial role in addressing the challenge of limited data availability. The primary aim of data augmentation is to increase the quantity and diversity of data, which in turn enhances a model's generalization capabilities. Before delving into the more recent advancements in data augmentation, it's essential to understand the classical methods that laid the groundwork in this domain. A visual summary of these classical data augmentation techniques is provided in Fig. 2.7.

One of the foundational approaches in data augmentation, as outlined by Xu et al. [52], is the iterative annotation of unlabeled web data to transform it into a strongly labeled dataset. This method is particularly significant when handling web data characterized by noisy and weak supervision tags. The process begins with training a part detector on a strong dataset and then applying it to web data for noise reduction and part annotation, eventually integrating the refined web data back into the original dataset.

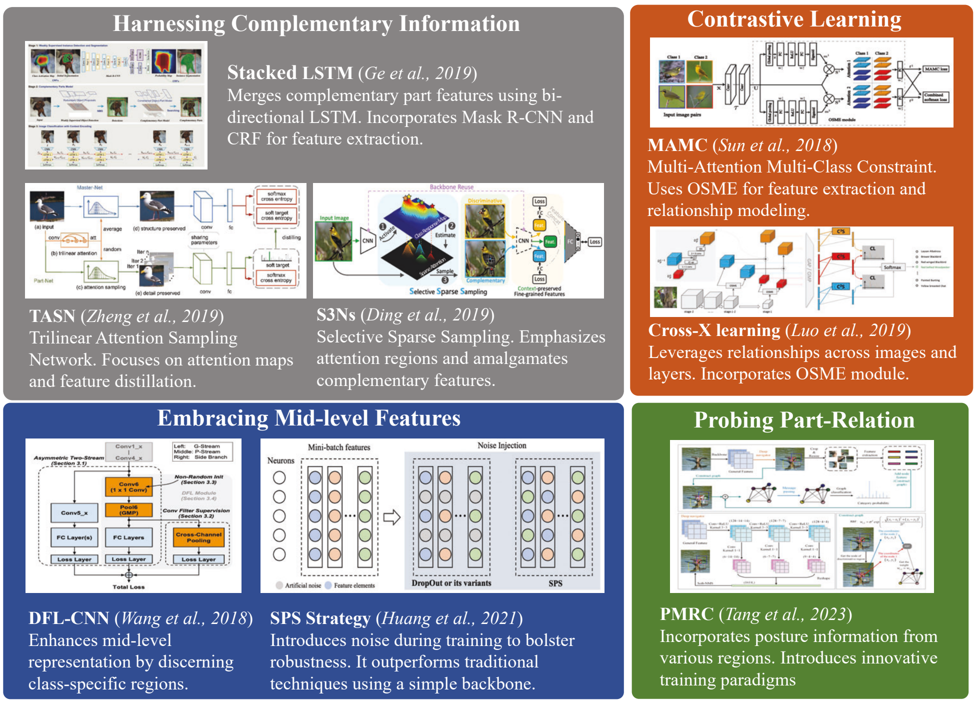
Shifting the focus to data transformation techniques, various strategies have been developed to create new data samples through image blending. An example is the Intra-class part swapping method introduced by Zhang et al. [53], which overcomes the limitations of standard augmentation techniques like Cutout, Mixup, and Cutmix. This method involves swapping parts between images of the same class, guided by attention maps, and employs affine transformations for compatibility in varying part sizes.

Another innovative approach is SnapMix by Huang et al. [54], which addresses the label noise issues commonly associated with mixed data methods. SnapMix uses semantic percentage maps derived from classification activation maps to accurately blend images and assign labels, minimizing the noise typically introduced by other mixing-based techniques.

2.1.6 Enhanced Representation Learning

Moving beyond the realms of data augmentation and basic feature extraction, we encounter advanced techniques in enhanced representation learning, crucial for fine-grained object recognition. These methods not only leverage expanded training data and diversity but also introduce increased task complexity to fortify the model's learning capabilities. An overview of these approaches is captured in Fig. 2.8.

A prime example of such an advanced methodology is the DCL (Destruction and Construction Learning), developed by Chen et al. [18]. DCL innovatively reconfigures images into Jigsaw puzzles before classification, involving a complex framework with four key components: the Region Confusion Mechanism (RCM) for generating Jigsaw images, a Classification Network for feature learning, an Adversarial Learning Network to address noise in Jigsaw features, and a Region Alignment Network focused on reconstructing the Jigsaw images. This method significantly enhances the accuracy of fine-grained object recognition by compelling the model to discern and interpret local features within the Jigsaw images, thus improving its ability to understand the interrelations of various image segments.



Harnessing Complementary Information

Stacked LSTM (*Ge et al., 2019*)
Merges complementary part features using bi-directional LSTM. Incorporates Mask R-CNN and CRF for feature extraction.

TASN (*Zheng et al., 2019*)
Trilinear Attention Sampling Network. Focuses on attention maps and feature distillation.

S3Ns (*Ding et al., 2019*)
Selective Sparse Sampling. Emphasizes attention regions and amalgamates complementary features.

Contrastive Learning

MAMC (*Sun et al., 2018*)
Multi-Attention Multi-Class Constraint. Uses OSME for feature extraction and relationship modeling.

Cross-X learning (*Luo et al., 2019*)
Leverages relationships across images and layers. Incorporates OSME module.

Embracing Mid-level Features

DFL-CNN (*Wang et al., 2018*)
Enhances mid-level representation by discerning class-specific regions.

SPS Strategy (*Huang et al., 2021*)
Introduces noise during training to bolster robustness. It outperforms traditional techniques using a simple backbone.

Probing Part-Relation

PMRC (*Tang et al., 2023*)
Incorporates posture information from various regions. Introduces innovative training paradigms

Figure 2.8: Enhanced Representation Learning in Fine-grained Object Recognition.

As fine-grained recognition tasks necessitate highly detailed object representations, modern approaches aim to accumulate rich, comprehensive feature sets. Earlier methods often struggled to capture enough detail, focusing on limited regions during the feature extraction process. However, contemporary leading strategies emphasize the acquisition of extensive representations, thus achieving peak accuracy in fine-grained recognition. These advanced techniques typically involve exploiting complementary informational regions, incorporating mid-level features, and utilizing contrastive learning methods. The next parts will delve into a detailed analysis of these enhanced representation learning methodologies.

Harnessing Complementary Information. Gleaning features from regions rich in complementary information invariably augments the robustness of feature representation. Ge et al. [55] introduced the Stacked LSTM approach, which accumulates complementary component features and utilizes a bi-directional LSTM (Long Short-Term Memory) network for their integration. The methodology encompasses three stages: initial instance detection and segmentation via Mask R-CNN [69] and CRF [70], followed by the extraction of several complementary regions, and lastly, the synthesis of these regions' features using a bi-directional LSTM for classification. Notwithstanding its remarkable results, its intricate integration of modules like Mask R-CNN amplifies the complexity of the model.

In contrast, Zheng et al. [71] proposed the streamlined TASN (Trilinear Attention Sampling Network) framework, which comprises three modules: trilinear attention, an attention-driven sampler, and a feature distiller. The method distills the essence of images into a series of attention maps, amplifies specific regions, and then transfers knowledge between models.

Drawing parallels to TASN, the Selective Sparse Sampling Network (S3Ns) by Ding et al. [72] also emphasizes attention-driven regions. However, S3Ns uniquely assemble sparse responses and amalgamate an array of complementary features, enriching the final representation.

Contrastive Learning. Integrating contrastive learning to discern relationships between features has proven instrumental in refining feature representation. Sun et al. [56] unveiled MAMC, encapsulating both an OSME (One-Squeeze Multi-Excitation) module for feature extraction and a MAMC (Multi-Attention Multi-Class) constraint to guide network learning. The methodology orchestrates relationships between diverse parts of an object, ensuring coherent and distinctive representations.

In a similar vein, Luo et al. [57] introduced Cross-X learning. This approach leverages relationships across images and layers to amplify the network's learning potential, ensuring that each layer captures unique information and mitigates redundancy.

Embracing Mid-level Features. Pioneering efforts have delved into leveraging deep mid-level features to accentuate subtle distinctions, showing promise in efficient fine-grained recognition. The rationale? Intermediate CNN layers inherently encode mid-level patterns, the supplementary mid-level feature extraction module remains lightweight, and the mid-level model shares computational processes with its high-level counterpart. However, direct mid-level feature acquisition often proves inadequate. Addressing this, Wang et al. [34] introduced a method to refine mid-level representation by discerning class-specific regions.

Moreover, Huang et al. [19] highlighted an inherent challenge in mid-level representation: the model's sensitivity to noise patterns. They proposed the SPS strategy, which introduces noise during training to mitigate the overt reliance on specific patterns, thus strengthening robustness and outperforming traditional techniques.

Probing Part-Relation. In fine-grained recognition, part-relation delves into the intricate relationships between distinct object parts. By understanding these dynamics, models can distinguish minute disparities in similar entities, refining classification in detailed tasks. For instance, Tang et al. [58] showcased PMRC, an innovative framework that amalgamates information from various regions to support classification, focusing on posture information and novel training paradigms.

2.1.7 Transformer-based Approaches

In the dynamic field of Fine-Grained Visual Categorization (FGVC), transformer-based techniques have been pivotal, evolving into two primary categories: 'Part Token Selection' and 'Part Correlation Learning'. These methods excel in discerning intricate relationships within images, especially valuable in scenarios with sparse labeled data.

Part Token Selection Methods. This approach focuses on selectively emphasizing specific parts or tokens of an image crucial for fine-grained classification. The cornerstone of this method is the Vision Transformer (ViT) framework, exemplified by the TransFG model [59]. TransFG revolutionized FGVC by adeptly selecting and emphasizing image parts critical to distinguishing nuanced categories. The effectiveness of this approach hinges on parsing an image into a sequence of patches (tokens) and using self-attention mechanisms to weigh these patches based on their relevance.

Subsequent advances such as FFVT [60] and Rams-trans [61] introduced innovations such as feature fusion and region-based attention. These enhancements allow for a more refined selection and weighting of image parts, leading to improvements in classification accuracy and computational efficiency.

Part Correlation Learning Methods. Contrastingly, part correlation learning focuses on understanding and representing the relationships between different parts of an object. This approach recognizes the criticality of the interplay between various image parts for accurate classification.

A notable development in this category is ViT-NeT [62], which integrates ViTs with a Neural Tree (NeT). This combination enhances the model's ability to learn the hierarchical and spatial relationships among parts, resulting in a more holistic image understanding, higher classification accuracy, and improved interpretability.

Another significant advancement is IELT [63], addressing limitations in ViT's multi-head self-attention mechanism. IELT's Multi-Head Voting, Cross-Layer Refinement, and Dynamic Selection modules synergize to refine the representation of features from different image parts. This approach ensures a more consistent ensemble learning process, effectively capturing the complex dynamics essential for fine-grained categorization.

In summary, the evolution of transformer-based methods in FGVC, marked by these two distinct approaches, is setting new benchmarks. Part Token Selection methods excel in identifying and emphasizing the most relevant image parts, while Part Correlation Learning methods focus on understanding and representing complex part relationships. Together, they are redefining accuracy and sophistication in computer vision research.

2.2 Datasets and Evaluation

The datasets and evaluation metrics are introduced here to contextualize the comparison across surveyed methods, and not as part of our experimental design, which is detailed in later chapters.

The datasets used for fine-grained image classification are summarized in Table 2.2, which provides an overview of various datasets categorized by their meta-class, number of categories, total images, and the distribution of training, validation, and testing sets. These datasets cover a wide range of fine-grained recognition tasks, including birds (CUB200-2011, NABirds), cars (Stanford Cars, CompCars), flowers (Oxford 102 Flower), food (Food-101), and other domains such as aircrafts, plants, fruits, and fungi. The diversity and scale of these datasets play a crucial role in advancing fine-grained classification research.

CUB-200-2011 [24] is a representative fine-grained recognition dataset with 11,788 images of 200 bird species, divided into 5,994 training and 5,794 testing images. Annotations include subcategory labels, key points, binary attributes, and bounding boxes.

FGVC-Aircraft [73] comprises 10,000 images of 100 variants of aircraft models, divided into 6,667 training images and 3,333 testing images. Each image includes a bounding box and four hierarchical labels.

Stanford Cars [74] contains 16,185 images of 196 car models, divided into 8,144 training images and 8,041 testing images.

Stanford Dogs [75] features 20,580 images of 120 dog breeds, with bounding boxes and breed labels. Each class has around 100 training images and at least 50 testing images.

NABirds [25] is a large-scale dataset with 48,562 images of 555 North American bird species. The dataset is split into 23,929 training images and 24,633 test images.

INat2017 [76] is a large-scale dataset with 13 base-level categories and 5,089 subordinary categories, totaling 858,170 images. It is divided into 579,184 training images, 95,986 validation images, and 183,000 test images.

Oxford 102 Flower [77] is an image classification dataset with 103 flower categories and 8,189 images.

Food-101 [78] comprises 101k images from 101 food categories, with 750 training images and 250 testing images per category. The dataset is also used for learning with the label noise evaluation.

DataSet	Meta-Class	Categories	Images	Training	Validation	Testing
CUB200-2011 [24]	Birds	200	11,788	5,994	-	5,794
FGVC-Aircraft [73]	Aricrafts	100	10,000	3,334	3,333	3,333
Stanford Cars [74]	Cars	196	16,185	8,144	-	8,041
Stanford Dogs [75]	Dogs	120	20,580	12,000	-	8,580
NABirds [25]	Birds	555	48,562	23,929	-	24,633
INat2017 [76]	Plants ect.	5,089	858,170	579,184	95,986	183,000
Oxford 102 Flower [77]	flowers	102	8,189	1,030	1,030	6,129
Food-101 [78]	Food	101	101,000	75,750	-	25,250
CompCars [79]	Cars	1716	51,304	25,652	-	25,652
VegFru [80]	Veg Friuts	292	160,731,	29,200	14,600	116,931
DF20 [81]	Danish Fungi	1604	295,938	248,466	-	27,608

Table 2.2: Summary of the fine-grained dataset

CompCars [79] is a car image dataset with 164,344 images from 1,716 models, including a fine-grained recognition subset of 51,304 images from 431 models. The subset is split into 30,955 whole car images and the rest as car parts, with half for training and half for testing.

VegFru [80] is a domain-specific dataset of 160,000 images from 292 vegetable and fruit categories, based on eating characteristics. Each category has at least 200 images, divided into 100 for training, 50 for validation, and the rest for testing.

DF20 [81] is the Danish Fungi 2020 dataset with 295,938 images from 1,604 fungi species. It provides accurate annotations, detailed metadata, and an unbalanced class distribution. The dataset avoids overlaps with ImageNet for better fine-tuning evaluation.

2.2.1 Dataset Biases and Generalizability in FGVC

Datasets play a crucial role in the development and evaluation of FGVC methods. However, biases inherent in these datasets can affect the generalizability of the developed models. Here, we outline some common biases and discuss their potential impact on FGIC methods:

Sampling Bias: Datasets may not represent the true distribution of the underlying population due to biases in the data collection process. This can lead to overfitting and poor generalization performance when models are tested on real-world data from different distributions.

Label Bias: Labeling errors or inconsistencies can introduce biases in the dataset. In the context of FGVC, annotating fine-grained categories often requires expert knowledge, which can be difficult to obtain. Errors or inconsistencies in labeling can result in models learning incorrect patterns or relationships between categories.

Imbalanced Data: Datasets may have imbalanced class distributions, with some fine-grained categories having significantly more samples than others. Imbalanced datasets can lead to biased models that perform poorly on underrepresented categories.

Viewpoint and Pose Bias: Datasets may have a bias towards certain viewpoints or poses, which can limit the model's ability to generalize to images captured from different perspectives.

Background and Context Bias: Images in the dataset may have specific backgrounds or contexts that are strongly correlated with certain fine-grained categories. Models trained on such datasets may rely on these correlations instead of learning the actual discriminative features of the categories.

2.2.2 Ethical and Privacy Concerns in FGVC

Addressing ethical and privacy concerns in FGVC is essential for responsible research and development. We briefly discuss these concerns and possible mitigation strategies:

Privacy Violation: Using images without consent, especially those containing identifiable information, can violate privacy. Researchers should obtain consent and anonymize identifiable information during the annotation process.

Data Bias: Biased datasets can result in unfair model behavior. Ensuring diverse data sources and evaluating performance across demographic groups can help identify and address biases.

Invasive Data Collection: Collecting images of sensitive subjects may cause harm. Adhering to guidelines and regulations can ensure non-invasive data collection that respects subjects and the environment.

Fairness of Annotation Work: The annotation process can be labor-intensive and require expertise. Providing fair compensation and working conditions for annotators is crucial for maintaining ethical standards in FGIC research.

2.2.3 Comparative Performance Analysis on Fine-Grained Datasets

In this chapter, we present a comprehensive comparison of various state-of-the-art techniques in the realm of fine-grained visual classification (FGVC). The comparison, as detailed in Table 2.3, encompasses a diverse array of methods evaluated across three prominent fine-grained datasets: CUB-200-2011, Stanford-Car, and FGVC-Aircraft.

Overview of Comparative Results. The Table2.3 succinctly summarizes the performance of various methods, highlighting their accuracy percentages on the specified datasets. Each method is characterized by its unique approach, the source of its foundational research, and the backbone network employed. This collection of data allows for a direct comparison of methodological effectiveness, offering insights into the evolution of techniques over time.

Analysis of Scalability and Computational Efficiency. To analyze the computational complexity of the reviewed algorithms, we obtained our ranking based on the backbone of the smallest supportable network for each method and the approximate GFLOPs of the backbone. The GFLOPs values we considered are as follows: VGG16 has approximately 15.4 GFLOPs, VGG19 has approximately 19.6 GFLOPs, ResNet50 has approximately 4.1 GFLOPs, ResNet101 has approximately 7.8 GFLOPs, DenseNet161 has approximately 7.8 GFLOPs, ViT - B - 16 has approximately 76.8 GFLOPs, Swin - Transformer - b has approximately 15.4 GFLOPs, and Inceptionv4 has approximately 3.4 GFLOPs. We classified the algorithms into three categories: low complexity (GFLOPs less than 10), medium complexity (GFLOPs greater than 10 and less than 30), and high complexity (GFLOPs greater than 30). We have added this computational complexity classification ranking in the backbone column of the table.

Regarding scalability, in addition to the computational complexity ranking, we have further analyzed the scalability of each method in the following ways. For low-complexity methods, such as those based on Inceptionv4, their relatively lower GFLOPs make them more scalable in scenarios where computational resources are limited. They can be more easily deployed on devices with lower processing power or in large-scale systems where multiple concurrent processes are required without overloading the system.

Medium-complexity methods, like those with backbones such as VGG19 and Swin-Transformer-B, while having higher computational requirements than the low-complexity ones, still offer a balance in scalability. They can handle moderately large datasets and are suitable for applications where a trade-off between accuracy and computational resources is needed. However, as the data size grows significantly, their scalability might be limited due to the increase in computational time and resource consumption.

High-complexity methods, for example, those using ViT-B-16, present challenges in scalability. Although they might offer superior performance in some cases, their high

GFLOPs values imply that they require substantial computational resources. In real-world applications with large-scale data or resource-constrained environments, their deployment might not be feasible without significant hardware upgrades or optimization. This could lead to longer processing times and potential bottlenecks in the system.

Trends and Observations. Several notable trends can be discerned from Table 2.3. There is a clear progression in the complexity and capability of backbone networks used, from earlier models like VGG and Inception to more advanced architectures like ResNet and Vision Transformers. This evolution reflects the continuous advances in neural network design and their impact on FGVC.

Another trend of increasing accuracy rates over time can be observed, indicating the progressive refinement of FGVC techniques. The introduction of transformer-based models, for example, marks a significant leap in performance.

The table also showcases a wide range of approaches, from spatial transformer networks (STN) to more recent innovations like P2P-Net and DCAL. This diversity underscores the dynamic nature of research in FGVC, with various methodologies being explored to tackle the intricate challenges of the field.

Finally, the recent surge in accuracy, particularly with transformer-based models, can be partly attributed to the adoption of self-supervised learning paradigms (SSL). SSL is a learning paradigm that takes advantage of the structure of the input data to learn useful representations without relying on explicit labels. This approach has gained popularity as it can alleviate the need for large-scale labeled datasets, which are often difficult to obtain in fine-grained categorization tasks. For example, training a model to predict the rotation angle applied to an input image can encourage the learning of rotation-invariant features, which can be beneficial to FGIC [82]. In addition, contrastive learning approaches can learn representations that capture the subtle differences between fine-grained categories [83].

2.2.4 Performance Analysis on Large-Scale Fine-Grained Datasets

This chapter offers a detailed evaluation of diverse methodologies applied to large-scale fine-grained datasets, specifically NABirds and INat2017. The comparative analysis is encapsulated in Table 2.4, which presents the accuracy percentages achieved by various state-of-the-art methods.

Method	Source	Backbone	Accuracy (%)		
			CUB-200-2011	Stanford-Car	FGVC-Aircraft
STN [84]	NIPS15	4xInception-v2 (M)	84.1	-	
B-CNN [48]	ICCV15	2xVGG16 (H)	84.1	91.3	84.1
Compact B-CNN [49]	CVPR16	1xVGG-16 (M)	84.0	-	-
RA-CNN [14]	CVPR17	3xVGG-19 (H)	85.3	92.5	88.2
Kernel-Pooling [85]	CVPR17	1xVGG-16 (M)	86.2	92.4	86.9
Low-rank B-CNN [13]	CVPR17	1xVGG-16 (M)	84.2	90.9	87.3
G2DeNet [86]	CVPR17	1xVGG-16 (M)	87.1	92.5	89.0
IBP-CNN [87]	BMVC17	1xVGG-19 (M)	85.8	92.0	88.5
Kernel-Activation [50]	ICCV17	1xVGG-16 (M)	85.3	91.7	88.3
MA-CNN [46]	ICCV17	3xVGG-19 (H)	86.5	91.5	89.9
GP-256 [88]	ECCV18	1xVGG-16 (M)	85.8	92.8	89.8
MaxEnt [89]	NIPS18	1xDenseNet161 (L)	86.5	93.0	89.7
PC [90]	ECCV18	1xDenseNet161 (L)	86.9	92.9	89.2
MAMC [56]	ECCV18	1xResnet-50 (L)	86.5	93.0	-
NTS-Net [91]	ECCV18	3xResnet-50 (M)	87.5	93.9	91.4
DFL-CNN [34]	CVPR18	1xResnet-50 (L)	87.4	93.1	91.7
iSQRT-CONV [92]	CVPR18	1xResnet-101 (L)	88.7	93.3	91.4
DCL [18]	CVPR19	1xResnet-50 (L)	87.8	94.5	93.0
Stacked LSTM (+Mask RCNN[69]) [55]	CVPR19	9xGoogleNet (M)	90.3	-	-
TASN [71]	CVPR19	1xResnet-50 (L)	87.9	93.8	-
Cross-X [57]	ICCV19	1xResnet-50 (L)	87.7	94.6	92.6
S3N [72]	ICCV19	3xResnet-50 (M)	88.5	94.7	92.8
MGE-CNN [15]	ICCV19	3xResnet-50 (M)	88.5	93.9	-
DF-GMM [93]	CVPR20	5xResnet-50 (M)	88.8	94.8	-
ACNet [94]	CVPR20	1xResnet50 (L)	88.1	94.6	92.4
SPS [19]	ICCV21	1xResnet-50 (L)	88.7	94.9	92.7
P2P-Net [47]	CVPR22	Resnet-50 (L)	90.2	95.4	94.2

TABLE 2.3: Comparison of different techniques on fine-grained datasets. Here, $n\times$ backbone means the method requires n forward pass of the backbone network in testing, while $\frac{4}{5}$xResnet-50 indicates that only the first four of five Conv blocks of Resnet-50 are needed. SPS* denotes results obtained using two mid-level branches.

Comparative Analysis. The table highlights the performance of several advanced techniques, each employing different backbone networks and developed from various research sources. The comparison spans two challenging datasets, NABirds and INat2017, known for their extensive variety of categories.

From Table 2.4, we observe the following trends: Advancement in Backbone Architectures: The table shows a progression from VGG16 to more advanced DenseNet and ResNet architectures. This evolution signifies continued improvements in backbone networks, adapting to the complexities of large-scale fine-grained datasets.

Diverse Methodological Approaches: Techniques range from B-CNN, which focuses on bilinear convolutional features, to more recent methods like SPS, which utilize stochastic

Method	Source	Backbone	Accuracy(%)	
			NABirds	INat2017
B-CNN [48]	ICCV15	2xVGG16	80.9	
MaxEnt [89]	NIPS18	1xDenseNet161	83.0	-
PC [90]	ECCV18	1xDenseNet161	82.8	-
SSN [95]	ECCV18	1xResnet101	-	65.2
TASN [71]	CVPR19	1xResnet101	-	68.2
MGE-CNN [15]	ICCV19	3xResnet101	88.0	-
Cross-X [57]	ICCV19	1xResnet-50	86.2	-
API-Net [96]	AAAI20	1xResnet-50	86.2	-
MRDNN [35]	TNNLS21	1xResnet50	-	70.5
SPS [19]	ICCV21	1xResnet-50	87.1	-
		1xResnet-101	87.9	-

Table 2.4: Comparison of different techniques on four fine-grained datasets. Here,$n \times$ *backbone* means the method requires n forward pass of the backbone network in testing, while $\frac{4}{5}$xResnet-50 indicates the first four of five Conv blocks of Resnet-50 is needed. SPS* denote results obtained by using two mid-level branches.

pooling strategies. This diversity highlights the multifaceted nature of research efforts in addressing the unique challenges of large-scale FGVC.

Improvements in Classification Accuracy: The methods listed demonstrate significant advances in accuracy, particularly in the NABirds dataset. For instance, MGE-CNN and Cross-X exhibit notable performance, indicating the effectiveness of these methods in handling the intricacies of large-scale datasets.

Challenges in INat2017: The results of the INat2017 dataset, although limited in the table, underscore the challenges posed by this dataset. Techniques like SSN and TASN, while achieving modest accuracy, highlight the ongoing need for innovative approaches to improve performance on such extensive datasets.

The analysis presented in Table 2.3 and Table 2.4 not only serves as a quantitative assessment of the methodological performance but also provides qualitative information on the evolving strategies used in large-scale FGVC. As the field progresses, further innovations are expected to emerge, pushing the limits of accuracy and computational efficiency in fine-grained visual classification on a grand scale.

Chapter 3

Learning multi-level weight-centric features for few-shot learning

3.1 Introduction

Despite the remarkable success achieved in visual recognition tasks [19, 68, 97, 98], deep learning models generally lack versatility and extendability, hindering their applicability in practice. For example, as data-hungry to learn massive parameters, deep neural networks often fail to work well in data-scarce environments [99, 100]. Besides, a trained model's prediction domain is usually not expandable unless re-executing the training process. In response to these deficiencies, there have been increasing efforts devoted to few-shot learning (FSL) [101–105]. Moreover, FSL exploration is gradually expanding in various research problems such as FKP recognition [106], medical image classification [107], object detection [108], text classification [109], and instance credibility inference [110].

FSL refers to a technique that takes advantage of the knowledge of base class data (provided auxiliary training set) to allow models to understand new concepts from only a few examples [109, 111–114]. Existing approaches to this problem mainly consist of meta-learning and weight-generation based frameworks. The former focuses on learning a meta-learner from base-class data to facilitate learning a new-task learner. Although meta-learning approaches achieve great success, they often require sophisticated training procedures and are difficult to extend to generalized few-shot learning (GFSL) settings. In contrast, the weight-generation framework delivers a more straightforward and flexible solution. This type of method first learns an embedding space from base-class data

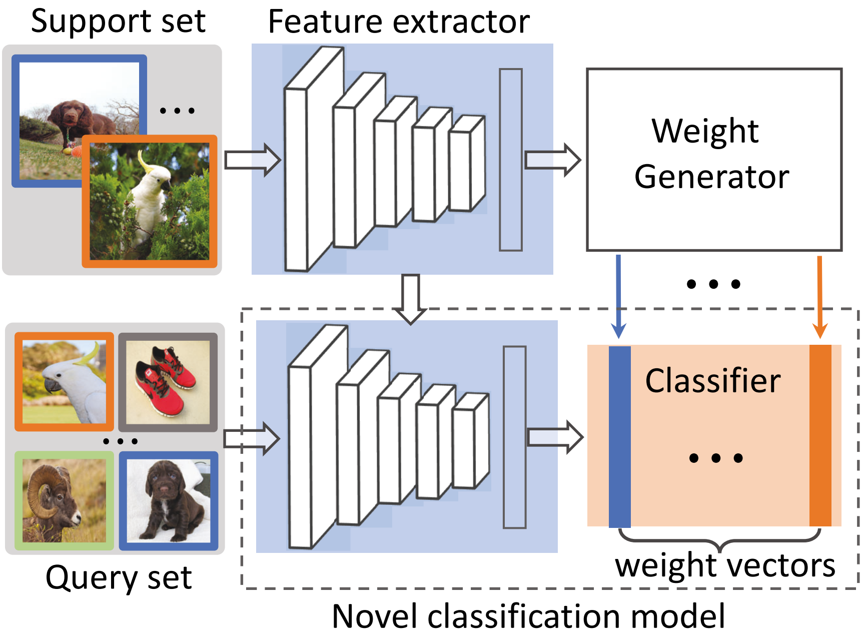


FIGURE 3.1: A general framework of weight generation methods for few-shot learning. Learning a good feature extractor plays a vital role in this framework, as it is used for novel models to extract image features and generate classifier weights for new categories.

and then utilizes the embedding of a support set (novel-class training samples) to construct the corresponding classifier weights(as illustrated in Fig. 3.1. This learning regime simplifies the few-shot learning problem by mainly focusing on feature learning and weight generation, enabling a trained model's extendability.

In the framework, feature learning is crucial due to its dual-use mechanism (representing images and constructing classifiers). However, existing methods learn a feature extractor without considering three essential issues associated with its dual functionality: representation transferability, base-class memorability, and prototype-ability. Transferability refers to whether the learned representation from base-class data is transferable to novel-class data. Recent works [101, 115–117] mainly extract features from the last Conv layer of deep models, leading to a lower transferability of the feature extractor. This can be attributed to the fact that higher layer activations with higher specialization to base-class tasks are less transferable to novel-class tasks [118] when there is a large domain gap between the two tasks. The memorability and prototype-ability are more related to the quality of the generated classifier weights. A GFSL model requires the preservation of the classification performance for the base-class data. This requirement necessitates base-class memorability to prevent novel-class weights from classifying base-class data to novel classes. The prototype-ability refers to the feature extractor's capacity to allow few-shot examples to approximate their class-specific prototype. Current methods attempt to complement these two capabilities by learning a weight-generation network.

Nevertheless, similar to meta-learning approaches, they need to train a new task-specific learner for weight generation, limiting the flexibility to construct few-shot classification models. In addition, the weight generator learns the required information from the extracted features but does not access more information through the feature learning stage. In summary, the existing weight-generation methods do not fully consider the feature extractor's dual-capacity in FSL, which may be a bottleneck of performance.

In this chapter, we propose a multi-level weight-centric feature extractor to complement the capacity of current weight-generation methods. We first introduce a weight-centric training strategy to increase the possibility that each sample can approximate its category prototype. Specifically, we fix the classifier weights in the latter learning stage and then enforce samples closer to their corresponding classifier weight in the embedding space. Besides, we build the multi-level feature by incorporating a mid-level and relation-level learning branch with high-level feature learning. The mid-level learning branch extracts mid-level features from intermediate layers, while the relation-level one obtains category-relation information from softening predictions. We finally integrate the multi-level information extraction and the weight-centric strategy into an overall feature learning framework.

Our proposed method ensures the comprehensiveness of the feature extractor to advance few-shot learning, which can be understood in the broader context of the seven paradigms outlined in Chapter 2. Specifically, it integrates elements from multiple paradigms, including Enhanced Representation Learning and Transformer-based Approaches, while also sharing characteristics with the Data Augmentation paradigm through its episodic training setup and support-query structure. On the one hand, the weight-centric strategy reduces the intra-class variance, improving feature representation generalization. It also pushes data points that are closer to the hyperplane far away. This effect indirectly achieves larger margin classification boundaries, increasing the feasibility of constructing a discriminative decision boundary based on a few samples. On the other hand, mid-level features are more transferable [119] to novel classes, and the relation-level representation exhibits higher-level abstraction and is more specific to base categories. Therefore, by jointly representing images using these two additional information sources, the resulting model has higher transferability for characterizing novel classes and better preserves the classification capability for base classes.

We extensively evaluate our approach on two low-shot classification benchmarks in both standard and generalized FSL learning settings. Experiments show that our proposed method significantly outperforms its counterparts in both learning settings and using different network backbones. We also demonstrate that the mid-level features exhibit strong transferability even in a cross-task environment, and the relation-level features help preserve base-class accuracy in the generalized FSL setting.

The major technical innovations introduced in this chapter include:

- We propose a weight-centric learning strategy that helps reduce the intra-class variance of novel-class data.
- We propose a multi-level feature learning framework, which demonstrates its strong prototype-ability and transferability even in a cross-task environment for few-shot learning.
- We extensively evaluate our approach on two low-shot classification benchmarks in both standard and generalized FSL learning settings. Our results show that mid-level features exhibit strong transferability even in a cross-task environment while relation-level features help preserve base-class accuracy in the generalized FSL setting.

3.2 Related Works

Recently proposed approaches to the few-shot learning problem can be roughly divided into meta-learning based[120–125] and weight-generation based approaches[115–117, 126, 127].

3.2.1 Meta-learning based methods

Meta-learning based methods tackle the few-shot learning problem by training a meta-learner to help a learner to effectively learn a new task on very few training data[121, 122, 128–130]. Most of these approaches are normally designed based on some standard practices for training deep models on limited data, such as finding good initialization of weights[121] or performing data augmentation[122] to prevent overfitting. For example, Finn et al. [121] propose to learn a set of parameters to initialize the learner model so that it can be quickly adapted to a new task with only a few gradient descent steps; Hariharan and Girshick [122] deal with data deficiency in a more straightforward way, in which a generator is trained on meta-training data and used to augment feature of novel examples for training the learner. Another line of work addresses the problem in a ”learning-to-optimize” way[123, 129]. For example, Ravi and Larochelle [123] trained an LSTM-based meta-learner as an optimizer to update the learner and store the previous update records in the external memory. Although this group of methods achieves promising results, they require either the design of complex inference mechanisms[131] or the further training of a classifier for novel concepts[121, 123]. Our work focuses on learning a feature extractor with dual functions(i.e. feature representation and classifier weight generation) for FSL problems. Therefore, the major difference from meta-learning techniques is that our method only needs to learn a base model and can construct new models directly using sample features.

3.2.2 Weight-generation based methods

Weight-generation based methods mainly learn an embedding space in which images are easy to classify using a distance-based classifier such as cosine similarity or nearest neighbor. To do so, Koch et al. [126] trained a Siamese network that learns a metric space to perform comparisons between images. Vinyals et al. [115] proposed Matching Networks to learn a contextual embedding, with which the label of a test example can be predicted by looking for its nearest neighbors from the support set. The prototypical networks[132] determine the class label of a test example by measuring the distance from all the class means of the support set. Since the distance functions of these two works are predefined, [133] further introduced a learnable distance metric to compare query and support samples. Ji et al. [134] proposed a re-weighting mechanism to improve the instance representativeness and an information-guidance mechanism to encode discriminative knowledge. Guo and Cheung [135] presented an Attentive Weights Generation via an Information Maximization strategy that generates optimal classification weights for the query sample within the task by self-attention and cross-attention paths.

The most related methods to ours are [116, 117, 136]. These approaches learn a feature representation by cosine softmax loss, allowing a few novel examples to construct the classifier. Our proposed method differs from them in two folds. First, they only learn a single level of representation, resulting in a limited representation capability, while ours constructs a multi-level model that considers multiple knowledge sources. Furthermore, those methods do not explicitly consider prototypeability(the ability to approximate the corresponding prototype by one or several sample features) in learning the feature extractor. In contrast, we introduce a weight-centric learning strategy that makes it more feasible to construct classifier weights from a few samples.

3.2.3 Analyzing the transferability of ConvNets.

Deep learning models are quite data-hungry but nonetheless, transfer learning has been proven highly effective in avoiding over-fitting when training larger models on smaller datasets[30, 137, 138]. These findings raise interest in studying the transferability of deep model features in recent years. Yosinski et al. [118] experimentally show how transferable each layer is by quantifying the generality versus specificity of its features from a deep ConvNet, and suggest that higher layer activations with higher specialization to source tasks are less transferable to target tasks. Pulkit et al. [139] investigate several aspects that impact the performance of ConvNet models for object recognition. Hossein et al. [140] identify several factors that affect the transferability of ConvNet features and demonstrate that optimizing these factors helps to transfer tasks. However, these works mainly explore the transferability and generalization ability of ConvNet features in terms of target datasets where the training samples are much more than in the few-shot setting. In this work, we investigate the capacities of the intermediate layer, the last feature layer, and the softmax logits to perform a few-shot learning tasks.

3.3 Methodology

In this chapter, we first introduce some general notation used throughout the thesis. We first briefly review a general weight-generation-based framework for few-shot learning. We further introduce our method for learning base models. Finally, we describe how to utilize these base models in few-shot learning.

3.3.1 Notation

Let $f_\theta(\cdot) \in \mathbb{R}^d$ be a feature extractor parameterized by Θ and $W \in \mathbb{R}^{d \times c}$ be a weight matrix of a linear classifier. Here, d is the dimension of the output feature and c is the number of labels for the classification task. We further define $M(\cdot)$ as a neural network classification model such that $M(f_\theta(x), W) = W^T f_\theta(x)$ given an input image x. We denote the training set D_{train} and the test set D_{test}. Slightly different from the general classification setting, few-shot learning trains a model on the training data that consists of a base- and novel-class dataset, that is $D_{train} = D^b_{train} \cup D^n_{train}$. Here $D^b_{train} = \{(x_i, y_i), y_i \in Y^b\}^{N^b}_{i=1}$ is an abundant dataset while $D^n_{train} = \{(x_i, y_i), y_i \in Y^n\}^{N^n}_{i=1}$ contains very few samples for each label; Y^b and Y^n refers to two different label spaces and $Y^b \cap Y^n = \emptyset$. We further denote the weight matrices W^b and W^n which correspond to Y^b and Y^n respectively.

3.3.2 Weight-generation-based framework

Weight-generation-based approaches have gained increasing attention in recent years, due to their simplicity and flexibility. The general framework for these methods usually consists of two stages: base-model learning and weight generation. As shown in Fig.3.2, this framework first learns a classification base-model on base-class dataset. In the second stage, based on the feature extractor $f_\Theta(\cdot)$ and classifier weights W^b of the base-model, a weight generator $g_\phi(.)$ is used to infer the weight vector w given training set $X^y = \{x_1^y, ..., x_k^y\}$. Here, the label y is in an unseen label space Y^n and k is usually a small number. In recent literature, there are two typical weight generators: average-based $w_{avg} = g^{avg}(f_\Theta(X^y))$ and attention-based $w_{att} = g_\phi^{att}(f_\Theta(X^y), W^b)$. The former simply computes the mean of the normalized features of training samples, which is expressed as:

$$w_{avg} = \frac{1}{k}\sum_{i=1}^{k} z_i, \tag{3.1}$$

where z_i is a L_2 norm of the feature vector $f_\Theta(x_i^y)$.

The second one employs an attention-based mechanism to exploit both the sample features and the base-class weights in generating the novel-class weights. The weight computation for an unseen label is expressed as:

$$w_{att} = \phi_{avg} \odot w_{avg} + \phi_{att} \odot (\frac{1}{k}\sum_{i=1}^{k}\sum_{b=1}^{K_b} Att(\phi_q z_i, k_b) \cdot w_b \tag{3.2}$$

where $odot$ is the Hadamard product,$\phi_{avg}, \phi_{att}, \phi_q$ are learnable parameters,$Att(.,.)$ is an attention kernel, and $\{k_b \in R^d\}_{b=1}^{K_b}$ is a set of K_b learnable keys.

3.3.3 Multi-level Weight-centric Representation Learning

Fig.3.3 provides an overview of our proposed method. The method mainly consists of two techniques: a multi-level feature extractor and a weight-centric feature learning strategy. The former aims to explicitly enforce each single sample feature vector closer to its corresponding classifier weight. Specifically, we construct three levels of feature representations namely mid-level, high-level and relation-level. The mid-level representation captures more subtle discriminative patterns, such as subordinary components of object parts, while the high-level encodes more holistic information. The relation-level is designed to describe the input's category structural relations, like how the input image relates to other categories. The second technique aims to obtain multiple representations that encode different levels of semantic information. Overall, the multi-level extractor improves the representation ability by considering multiple sources of information, and the weight-centric strategy increases the feasibility of generating classifier weights from

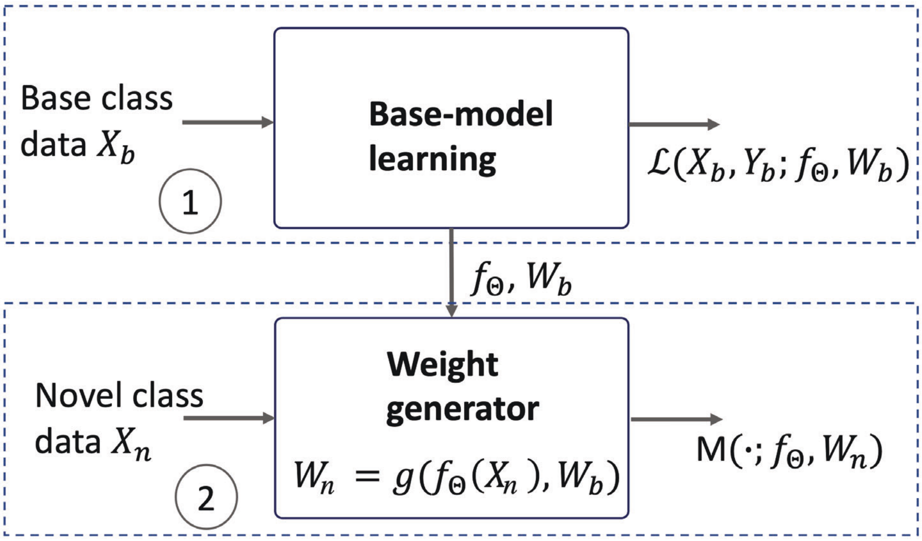

Base class
data X_b
1
Base-model
learning
$\mathcal{L}(X_b, Y_b; f_\Theta, W_b)$
f_Θ, W_b
Novel class
data X_n
2
Weight
generator
$W_n = g(f_\Theta(X_n), W_b)$
$M(\cdot; f_\Theta, W_n)$

FIGURE 3.2: A general weight-generation-based framework for few-shot learning. Here, $\mathcal{L}$ is the loss function for learning a base model on base-class data. f_Θ and W_b are the feature extractor and classifier weights of the base model. $g(\cdot)$ is a weight generator that can be defined or learned from data. $M(\cdot)$ is a novel model built for novel categories.

few-shot sample features. These two techniques can seamlessly join together to provide a simple and effective solution to a few-shot learning problem.

In this chapter, we first review cosine softmax loss for few-shot learning. We then introduce our proposed weight-centric embedding learning strategy. This strategy can be incorporated with cosine softmax loss to facilitate the subsequent step of generating weights from the few-shot training example.

Cosine Softmax Loss. In a standard classification framework, Softmax Loss is usually adopted for supervised learning. It generally refers to a Softmax Activation plus a Cross-Entropy Loss. Given an input (x_i, y_i), the softmax loss function is expressed as:

$$\ell_s(x_i, y_i) = -log(\frac{exp(w_{y_i}^T f_\Theta(x_i))}{\sum_j exp(w_j^T f_\Theta(x_j))}), \tag{3.3}$$

Where $f_\theta(\cdot)$ is the feature extractor and w_j is the j^{th} column of the weight matrix W of the classifier layer.

However, recent work shows that the softmax loss fails to learn a feature extractor that generalizes well to unseen categories[116, 117]. As discussed previously, the feature extractor of the base model is used to generate weights of novel categories. However, the

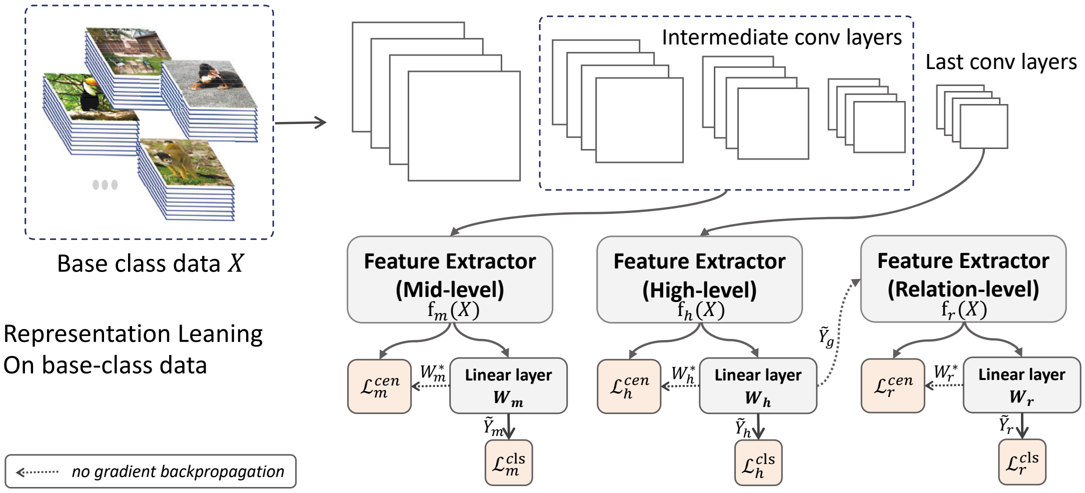


FIGURE 3.3: An overview of our learning framework for representation learning. The framework first constructs three levels of feature representations namely mid-level, high-level, and relation-level. For forwarding the networks, the outputs of the intermediate layer outputs are detached and fed to the mid-level feature extractor, the output of the last conv layer is forwarded to the high-level feature extractor, and the prediction logits of the high-level branch is detached and input and sent to the relation-level feature extractor. The three feature extractors are first trained to converge with the classification loss and then are further fine-tuned with both the classification and the weight-centric loss.

transferability gap of the model increases as the distance between tasks increases[118]. Therefore, the more significant between the base- and novel tasks, the poorer the performance of the few-shot learning model due to the weak transferability of the feature extractor. To ease this issue, [116, 117] proposed to adopt cosine softmax loss in learning the base model. Compared with softmax loss, Cosine softmax loss applies l_2-normalization on both the feature vector and the weight vector before the loss calculation, which is expressed as:

$$\tilde{w_j} = \frac{w_j}{\|w_j\|}, \tilde{f_\theta(x_i)} = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|} \tag{3.4}$$

This normalization step will cause the softmax function to fail producing a one-hot categorical distribution, making the neural networks hard to converge. As suggested in[116], a simple solution to this is to introduce a trainable scale factor s into the softmax function. Thus, the cosine softmax loss function is expressed as:

$$\ell_{cs}(x_i, y_i; \Theta, W) = -log(\frac{s \cdot exp(\tilde{w}_{y_i}^T \tilde{f}_\Theta(x_i))}{\sum_j exp(\tilde{w}_j^T \tilde{f}_\Theta(x_j))}), \tag{3.5}$$

Based on this loss function, [116, 117] learn the feature extractor by minimizing the cost function:

$$\mathcal{L}_{cs} = \frac{1}{N}\sum_{i}^{N}(\ell_{cs}(x_i, y_i; \Theta, W)) + \lambda R(W), \tag{3.6}$$

where $\lambda R(W)$ is a weight L_2 regularization term.

Weight-Centric feature learning. As illustrated in Fig.3.4 (a) and (b), learning with cosine softmax loss reduces intra-class variations by comparison with original softmax loss. Thus, it increases the feasibility of characterizing an unseen concept with few shot examples. [116, 117] assume that the samples of the same class are concentrated in the feature space learned with cosine softmax loss, then the feature embedding of some random samples can be used to approximate the classifier weights. However, this assumption is not strictly held in some cases, such as data with large intra-class variance and small inter-class variance might tend to be scattered in the feature space. To ensure that using one or a few embedded points of each category can construct a stable decision boundary, we explicitly constraint a feature point $\tilde{f}(x_i)$ should be near its classifier weight $\tilde{w}_{y_i}$, after the classifier is learned, and the constraint loss is given by

$$\ell_{cen}(x_i, w^*_{y_i}; \Theta) =\| f_{\Theta}(x_i) - w^*_{y_i} \|^2, \tag{3.7}$$

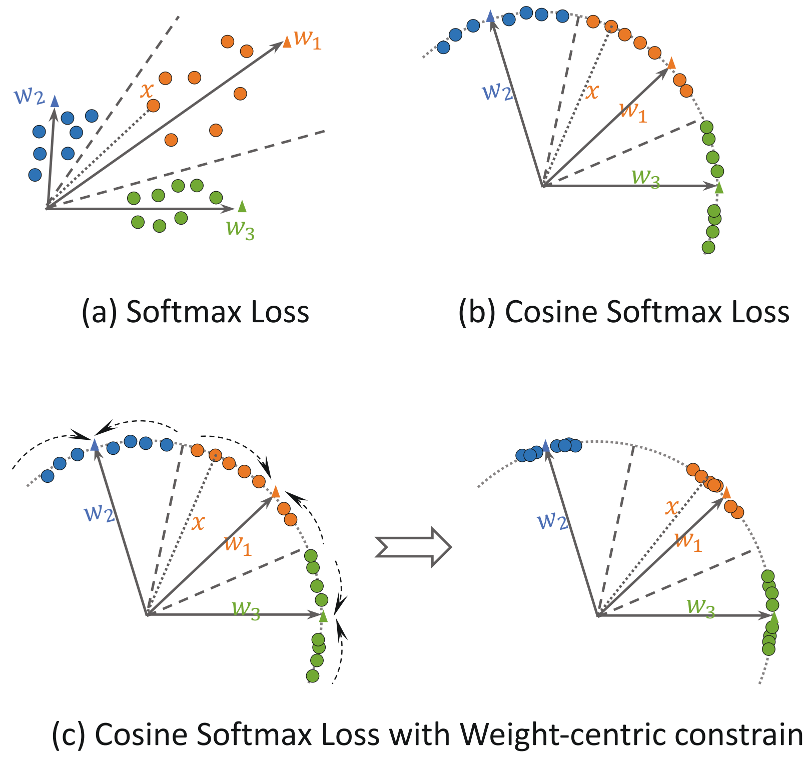
Where $w^*_{y_i}$ represents the sample x_i's corresponding class weight vector, which specifically refers to the y_i^{th} column of a constant matrix W^*. We obtain W^* from the classifier layer after first training the model to converge using the cost function ℓ_{cs}. To couple the constraint with cosine softmax loss, we also apply l_2-normalizaiton on both the feature vector and the weight vector. Thus, the weight-centric constraint can be rewritten as

$$\ell_{cen}(x_i, w^*_{y_i}; \Theta) =\| \frac{f_{\Theta}(x_i)}{\|f_{\Theta}(x_i)\|} - \frac{w^*_{y_i}}{\|w^*_{y_i}\|} \|^2 . \tag{3.8}$$

By integrating the cosine softmax loss and the weight-centric constraint, we now have the cost function $\mathcal{L}$.

$$\mathcal{L} = \begin{cases} \mathcal{L}_{cs} & \mathcal{L}_{cs} > \epsilon \\ \mathcal{L}_{cen} + \mathcal{L}_{cs}, & otherwise, \end{cases} \tag{3.9}$$

where $\mathcal{L}_{cen} = \frac{1}{N}\sum_i^N(\ell_{cen}(x_i, w^*_{y_i}))$ and $\mathcal{L}_{cs} > \epsilon$ means that the stopping criteria is not met when training with loss $\mathcal{L}_{cs}$. Since $\mathcal{L}_{cen}$ required W^* as input, we optimize the cost function using a two-stage algorithm which is detailed in Algorithm.**1**.



(a) Softmax Loss

(b) Cosine Softmax Loss

(c) Cosine Softmax Loss with Weight-centric constrain

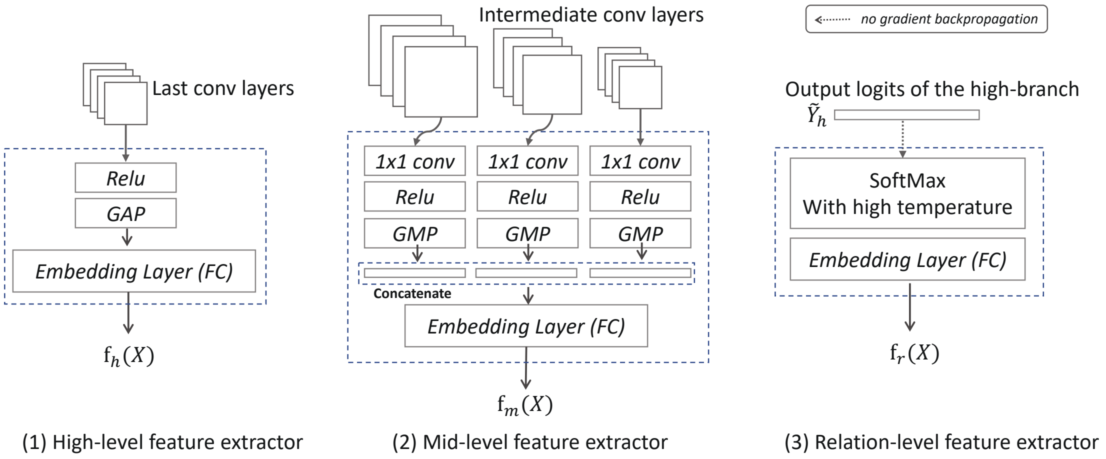
FIGURE 3.4: A geometry interpretation for learning feature space with different loss functions

As illustrated in Fig.3.4(c), the weight-centric constraint pushes samples closer to their corresponding classifier weights, which brings two advantages. First, it enforces the neural network to learn a feature space with smaller intra-class variance. Moreover, the constraint also implicitly drives samples far away from the decision boundary. This increases the feasibility of constructing a discriminative decision boundary based on a small number of samples.

A good representation of generalized few-shot learning is that it can generalize well to novel concepts while maximizing its original ability to discriminate base categories. A single high-level feature representation usually limits the capacity to meet these criteria simultaneously. In this chapter, we introduce two additional levels of representation named mid- and relation-level to complement the representative capacity of high-level representation.

High-level feature extractor is a common practice in most existing few-shot learning methods. As illustrated in Fig. 3.5 (1), it takes inputs from the last convolutional layer and then maps them into an embedding space after applying global-average pooling. This design results in the extraction of the features that naturally capture the global

```
Algorithm 1: Learning weight-centric features
Input  : Base-class Training data $\{X, Y\}$,feature extractor with parameters of $\Theta$
         ,linear classifier weights $\mathcal{W}$.
Output: Updated $\Theta$ and $\mathcal{W}$
Initialize parameters $\Theta$ and $\mathcal{W}$
while $\mathcal{L}_{cs}$ not converge do
    Sample a minibatch of $m$ examples from the training set $\{x^{(1)}, ..., x^{(m)}\}$ with
      corresponding targets $y^{(i)}$;
    Compute gradient: $g_\Theta \longleftarrow \frac{1}{m} \nabla_\Theta \sum_i \mathcal{L}_{cls}(x^{(i)}, y^{(i)}; \Theta, \mathcal{W})$ ;  ▷ $\mathcal{L}_{cs}$ is computed
      using eq.3.6
    Compute gradient: $g_\mathcal{W} \longleftarrow \frac{1}{m} \nabla_\mathcal{W} \sum_i \mathcal{L}_{cls}(x^{(i)}, y^{(i)}; \Theta, \mathcal{W})$ ;
    update $\Theta$ and $\mathcal{W}$ ;
end
$\mathcal{W}^* \longleftarrow \mathcal{W}$ ;                                ▷ Frozen classifier weights
while $\mathcal{L}_{centric}$ and $\mathcal{L}_{cs}$ not converge do
    Sample a minibatch of $m$ examples from the training set $\{x^{(1)}, ..., x^{(m)}\}$ with
      corresponding targets $y^{(i)}$ ;
    Compute gradient: $g_\Theta \longleftarrow \frac{1}{m} \nabla_\Theta \sum_i (\mathcal{L}_{cls}(x^{(i)}, y^{(i)}; \Theta, \mathcal{W}^*) + \mathcal{L}_{cen}(x^{(i)}; \Theta, \mathcal{W}^*))$;
    update $\Theta$;
end
```



(1) High-level feature extractor

(2) Mid-level feature extractor

(3) Relation-level feature extractor

FIGURE 3.5: Sub-network structures for different levels of feature extractor

visual discriminative patterns because of the high-level feature abstraction source and the property of the pooling operation.

Mid-Level feature extractor aims to obtain features that focus more on encoding mid-level discriminative patterns. Compared with high-level features, it exhibits better generalization ability to represent novel concepts, but weaker discriminatory power for the base concepts. This can be attributed to the fact that it tends to abstract information that is less specific to the base concepts. A naive scheme to learn mid-level features is to plug an additional global-extractor head on top of the intermediate layers. However, this solution might learn features more similar to the high-level ones because of the global average pooling operation, though the input source is switched to the lower layers. To avoid such undesirable effects, we designed the mid-level feature extractor, shown in Fig. 3.5 (2). Specifically, we insert a 1x1 Conv layer on top of each intermediate layer and employ global-max pooling to prevent the 1x1 Conv layer from learning global abstraction. Lastly, we concatenate all the intermediate-layer features into one and map it into embedding space to form a compact mid-level representation.

Relation-level feature extractor. As discussed previously, the model's generalization to novel concepts can be improved by incorporating the mid-level representation. However, its ability to classify base classes is degrading when the label space expands with more novel classes (some base-class examples might be misclassified to novel classes. Thus, we propose to preserve this ability by encoding more specific information of the base classes. Specifically, we introduce another relation-level representation that describes an input using its structural relationships within the base classes. This representation is more specific to base classes than the high- and mid-level representations. Although it has a poor generalization to novel concepts, it helps strengthen the classification capacity for base classes. As shown in Fig. 3.5 (3), the relation-level extractor takes inputs from the predicted $\tilde{Y}_h$ from the high-level branch. Here, when the $\tilde{Y}_h$ from a trained model is fed to a softmax layer, the outputs will tend to be a one-hot vector, which fails to describe the structural relation of the data over classes. Therefore, we feed $\tilde{Y}_h$ to a softmax function with a high temperature, so that it can encode a richer class structural information of the data. Finally, we use this softened prediction output to learn the embedding space that characterizes the similarity of samples according to their categorical distribution.

Jointly Learning multiple feature extractors. As shown in Fig.3.3, we learn the three feature extractors using three classification branches that are all based on a single network backbone. We also apply the weight-centric learning strategy for each branch. Thus, the overall classification loss and weight-centric loss

$$\begin{aligned} \mathcal{L}_{cs} &= \mathcal{L}_{cs}^{m} + \mathcal{L}_{cs}^{h} + \mathcal{L}_{cs}^{r}, \\ \mathcal{L}_{cen} &= \mathcal{L}_{cen}^{m} + \mathcal{L}_{cen}^{h} + \mathcal{L}_{cen}^{r} \end{aligned} \tag{3.10}$$

respectively. Finally, our overall cost function is obtained by substituting these two equations into Eq.3.9.

Model combination. After training the base models using our proposed method, we have three base models $M(f_m(x), W_m^b)$,$M(f_h(x), W_h^b)$,and $M(f_r(x), W_r^b)$, which denote the mid-, high-, and relation-level classification model respectively. We simply combine them into a single model $M(f_C(x), W_C^b)$ by concatenating their normalized features and classifier weights separately. Here, $f_C(x) = concat(\frac{f_m(x)}{\|f_m(x)\|}, \frac{f_h(x)}{\|f_h(x)\|}, \frac{f_r(x)}{\|f_r(x)\|})$ forms a multi-level feature extractor and $W_C^b = concat(W_m^b, W_h^b, W_t^b)$ is the classifier weight matrix for base categories. Given a test image x^b, this model can be used to predict the label in the base label space Y^b, that is $argmax(M(f_C(x), W_C^b)) \in Y^b$.

Generating weights for few-shot learning. Now, we can utilize the feature extractor $f_C(\cdot)$ and weight matrix W_C^b to construct different models for different few-shot learning settings. We first construct the weight matrix W_C^n for Y^n using a weight generator (AvgGen [116] or AttGen [117]). Then, we can build classification models $M(f_C(x), W_C^n)$ and $M(f_C(x), [W_C^b, W_C^n])$ for standard and generalized few-shot learning scenario respectively. Here, the weight matrix W_C^n is obtained by stacking each weight vector in order according to its label index in Y^n.

Let Y^b and Y^n denote the base- and novel-label space respectively, we obtain its corresponding weight vector w^y by normalizing the prototype of the given k training samples $\{x_1^y, ..., x_k^y\}$.

$$w^y = \frac{\frac{1}{k}\sum_{i=1}^n f(x_i^y)}{\|\frac{1}{k}\sum_{i=1}^n f(x_i^y)\|}, \tag{3.11}$$

where $f(.)$ is the multi-level feature extractor derived from the combined base model. Now, Given an unseen label space, we can build classification models $M(f(x), W^n)$ and $M(f(x), [W^b, W^n])$ for standard and generalized few-shot learning scenarios respectively. Here, the weight matrix W^n is obtained by stacking each weight vector in order according to its label index in Y^n, W^b is the weight matrix derived from the combined base model.

3.4 Experiments

3.4.1 Datasets and evaluation metrics

We validate our proposed method on Low-shot-ImageNet[122] and Low-shot-CUB[116] based on three performance metrics.

Low-shot-ImageNet contains 193 base categories, 300 novel categories, 196 base categories, and 311 novel categories, respectively. The first two groups are made for validating hyper-parameters, and the remaining two groups are used for the final evaluation.

Low-shot-CUB is constructed from the Caltech-UCSD bird dataset[24]. The dataset consists of 100 base classes and 100 novel classes. Since each category of this dataset contains only about 30 images, we repeated 20 experiments and took the average top-1 accuracy.

Performance evaluation metrics. Few-shot learning methods are evaluated differently according to different few-shot learning settings. These performance measures mainly differ in the way of constructing a test dataset. To evaluate our proposed method in both the standard and generalized setting, we use three evaluation metrics summarized below:

1) Novel/Novel: the model's performance is measured by the accuracy of novel test examples within the novel label space, that is, $D_{test} = \{(x_i, y_i) \in D^n_{test}, y_i \in Y^n\}$.

2) Novel/All: the model's performance is measured only by the accuracy of novel test examples in all label space, that is, $D_{test} = \{(x_i, y_i) \in D^n_{test}, y_i \in Y^b \cup Y^n\}$.

All: the performance of the model is measured by the accuracy of all the test examples in all the label spaces, that is, $D_{test} = \{(x_i, y_i) \in D^b_{test} \cup D^n_{test}, y_i \in Y^b \cup Y^n\}$.

Here, the standard few-shot learning setting only considers Novel/Novel as the major performance measure, while the generalized setting considers the results of both Novel-/All and ALL. We report the results of these metrics based on multiple tries. Specifically, in our experiments, we randomly select training images of the novel categories repeat experiments 100 times, and finally report the mean accuracies within 95% confidence intervals.

3.4.2 Network architecture and training details

Network architecture. We conduct experiments on the Few-shot-Imagenet benchmark using ResNet-10 and -50[7] architecture in our learning framework. For experiments on the Few-shot-CUB dataset, asQi et al. [116] obtained their results based on Inception V1[141], we implement our method based on the same network structure for performance comparison.

Training details. For all experiments on imageNet based few-shot benchmarks, we trained our method from scratch for 90 epochs on the base classes. The learning rate starts from 0.1 and is divided by 10 every 30 epochs with a fixed weight decay of 0.0001. We then further fine-tune the model with the classifier-centric constraint with a small learning rate 0.0001. For the CUB dataset experiment, all the pre-trained models we used are from the Pytorch official model zoo. During training, initial learning of 0.001 decreases by 0.1 at 30 epoch intervals.

3.4.3 Results and analysis

We evaluated the performance of the proposed method on two low-shot benchmarks.

	Novel / Novel					Novel / All					All				
Method	n=1	2	5	10	20	n=1	2	5	10	20	n=1	2	5	10	20
Pro. Nets [132] (from [139])	39.4	54.4	66.3	71.2	73.9	-	-	-	-	-	49.5	61.0	69.7	72.9	74.6
Log. Reg. (from [101])	38.4	51.1	64.8	71.6	76.6	-	-	-	-	-	40.8	49.9	64.2	71.9	76.9
Log. Reg w/G. (from [101])	40.7	50.8	62.0	69.3	76.5	-	-	-	-	-	52.2	59.4	67.6	72.8	76.9
Pro. Mat. Nets [101]	43.3	55.7	68.4	74.0	77.0	-	-	-	-	-	55.8	63.1	71.1	75.0	77.1
Pro. Mat. Nets w/G [101]	45.8	57.8	69.0	74.3	77.4	-	-	-	-	-	57.6	64.7	71.9	75.2	77.5
SGM w/G. [122]	-	-	-	-	-	32.8	46.4	61.7	69.7	73.8	54.3	62.1	71.3	75.8	78.1
Batch SGM [122]	-	-	-	-	-	23.0	42.4	61.9	69.9	74.5	49.3	60.5	71.4	75.8	78.5
Mat. Nets [115](from [101, 122])	43.6	54.0	66.0	72.5	76.9	41.3	51.3	62.1	67.8	71.8	54.4	61.0	69.0	73.7	76.5
Wei. Imprint* + AvgGen [116]	44.05	55.42	68.06	73.96	77.21	38.70	51.36	65.89	72.60	76.21	56.73	63.66	71.04	74.05	75.47
	±.21	±.16	±.09	±.07	±.05	±.21	±.17	±.09	±.07	±.05	±.13	±.10	±.06	±.04	±.03
AvgGen (with retraining) [117]	45.23	56.90	68.68	74.36	77.69	39.33	50.27	63.16	69.56	73.47	54.65	64.69	72.35	76.18	78.46
	±.25	±.16	±.09	±.06	±.06	±.25	±.16	±.11	±.07	±.06	±.15	±.10	±.06	±.04	±.04
AttGen [117]	46.02	57.51	69.16	74.84	78.81	40.79	51.51	63.77	70.07	74.02	58.16	65.21	72.72	76.65	78.74
	±.25	±.15	±.09	±.06	±.05	±.25	±.15	±.12	±.07	±.06	±.15	±.09	±.06	±.04	±.03
TRAML [142] + AttGen	48.1	59.2	70.3	76.4	79.4	-	-	-	-	-	59.2	66.2	73.6	77.3	80.2
MLWC + AvgGen	48.22	58.77	69.71	74.45	76.91	44.06	55.83	68.15	73.36	76.07	58.96	65.18	71.28	73.63	74.78
	±.12	±.09	±.05	±.03	±.02	±.12	±.09	±.05	±.04	±.02	±.07	±.05	±.03	±.02	±.02
MLWC* + AvgGen	49.09	59.66	70.26	74.72	77.04	45.56	57.12	68.85	73.73	76.24	59.37	65.48	71.36	73.63	74.72
	±.11	±.08	±.04	±.03	±.02	±.11	±.09	±.05	±.03	±.02	±.07	±.05	±.03	±.02	±.02
MLWC* + AttGen	**50.87**	**62.13**	**72.61**	**77.02**	**79.67**	**46.18**	**57.21**	**68.63**	**73.64**	**76.59**	**61.72**	**68.58**	**75.35**	**78.29**	**80.03**
	±.22	±.15	±.09	±.06	±.23	±.15	±.09	±.09	±.07	±.05	±.14	±.08	±.06	±.05	±.03

TABLE 3.1: Comparison of top-5 accuracy with the state-of-art methods using Resnet-10 on the Low-shot-ImageNet dataset. Best are bolded. * indicates that we get 5 random crops from each training example, then use the average feature as the weight of novel class.

Low-shot-ImageNet. Table3.1 and Table3.2 provide the comparative results of different techniques using two network backbones in the large-scale Few-shot ImageNet dataset. First, we can observe that some existing methods show a significant improvement in one evaluation metric but a minor improvement in another. In comparison, our approach consistently achieves the best results in all evaluation metrics. Specifically, using the same weight generator AttGen, our method significantly outperforms the current best model TRAML[142] in testing the precision of the classification of the novel class and all classes. In addition, without learning the weight generator, our proposed method also achieves a performance comparable to the current top-performing methods that require training a weight generator. For instance, compared to the TRAML method that needs to learn an attention-based weight generator, our approach obtains a similar performance using the mean feature as classifier weights. All these results indicate that our learned representation yields better generalization ability and versatility for FSL learning.

Low-shot-CUB. Since the existing method reported in this dataset is based on the Inception V1 network, we first evaluate our method with the same backbone network. Table3.3 shows the results of the performance comparison of different approaches. Our proposed method outperforms all comparing methods by a large margin in all evaluation metrics. For instance, our method achieves top-1 accuracies of 30.72% and 37.65% under the 1 and 2 shot settings respectively, the previous best results are 21.40% and 28.69%. To evaluate our method's effectiveness on this dataset when using different network architectures, we further use the Resnet-50 as the backbone for both the Imprinting and our method and compare their performance. Table3.4 shows the corresponding results and shows the superior performance of our method in low-shot learning.

	Novel / Novel					Novel / All					All				
Method	n=1	2	5	10	20	n=1	2	5	10	20	n=1	2	5	10	20
Mat. Nets [115]	53.5	63.5	72.7	77.4	81.2	-	-	-	-	-	64.9	71.0	77.0	80.2	82.7
ProNets [132]	49.6	64.0	74.4	78.1	80.0	-	-	-	-	-	61.4	71.4	78.0	80.0	81.1
ProMN w/G [101]	54.7	66.8	77.4	81.4	83.8	-	-	-	-	-	65.7	73.5	80.2	82.8	84.5
SGM w/G. [101]	-	-	-	-	-	45.1	58.8	72.7	79.1	82.6	63.6	71.5	80.0	**83.3**	**85.2**
MLWC +	57.12	68.28	77.77	81.80	83.72	53.48	65.05	76.59	80.95	83.07	67.49	73.36	79.87	81.98	82.95
AvgGen	±.20	±.14	±.07	±.07	±.04	±.23	±.13	±.08	±.08	±.04	±.14	±.08	±.05	±.05	±.02
MLWC*+	**57.97**	**69.08**	**78.19**	**81.99**	**83.80**	**54.82**	**66.93**	**77.12**	**81.22**	**83.16**	**68.01**	**74.72**	**79.98**	81.99	82.88
AvgGen	±.20	±.15	±.06	±.07	±.03	±.22	±.05	±.05	±.08	±.03	±.13	±.09	±.05	±.05	±.02

TABLE 3.2: Comparison of top-5 accuracy with the state-of-art methods using Resnet-50 on the Low-shot-ImageNet dataset.Best are bolded. * indicates that we get 5 random crops from each training example, then use the average feature as the weight of the novel class.

	Novel / Novel					Novel / All					All				
Method	n=1	2	5	10	20	n=1	2	5	10	20	n=1	2	5	10	20
Gen. + Cla.[122])	-	-	-	-	-	18.56	19.07	20.00	20.27	20.88	45.42	46.56	47.79	47.88	48.22
Mat. Nets [115]	-	-	-	-	-	13.45	14.75	16.65	18.18	25.77	41.71	43.15	44.46	45.65	48.63
Imprinting [116]	-	-	-	-	-	21.26	28.69	39.52	45.77	49.32	44.75	48.21	52.95	55.99	57.47
Imprinting* [116]	-	-	-	-	-	21.40	30.03	39.35	46.35	49.80	44.60	48.48	52.78	56.51	57.84
MLWC	32.35	39.78	49.47	54.67	57.37	30.72	37.65	48.17	53.56	56.45	49.80	53.41	57.87	**60.46**	**61.61**
MLWC*	**33.56**	**40.82**	**50.28**	**54.67**	**57.53**	**30.87**	**39.01**	**49.17**	**53.66**	**56.61**	**49.96**	**53.73**	**58.18**	60.30	61.60

TABLE 3.3: Comparison of top-1 accuracy with the state-of-art methods on the Few-shot-CUB dataset. * indicates that we get 5 random crops from each training example, then use the average feature as the weight of the novel class.

	Novel / Novel					Novel / All					All				
Method	n=1	2	5	10	20	n=1	2	5	10	20	n=1	2	5	10	20
Imprinting* [116]	32.15	40.48	52.41	57.93	61.72	26.24	35.79	49.31	55.31	59.38	52.43	56.83	62.89	65.53	67.27
MLWC	35.91	44.91	56.95	62.48	66.01	33.54	43.47	56.21	61.96	65.61	55.45	59.58	64.94	67.32	68.78
MLWC*	**36.96**	**45.53**	**57.43**	**63.03**	**66.35**	**34.91**	**44.21**	**56.81**	**62.52**	**65.96**	**55.60**	**59.66**	**65.02**	**67.46**	**68.89**

Table 3.4: Comparison of top-1 accuracy with the state-of-art methods on the Few-shot-CUB dataset. * indicates that we get 5 random crops from each training example, then use the average feature as the weight of the novel class.

	Novel classes from ImageNet					Novel classes from CUB2011				
Method	n=1	2	5	10	20	n=1	2	5	10	20
High-level(baseline)	51.56	63.67	74.78	79.68	82.45	30.55	40.76	53.68	60.79	65.54
High-level(baseline)+WC	54.24	65.71	75.75	81.33	82.80	35.92	47.67	61.92	69.35	73.42
Mid-level	51.59	63.80	75.57	80.60	83.21	35.99	**48.40**	**62.51**	**70.26**	**74.92**
Relation-level	48.94	58.64	69.23	73.32	75.65	24.45	32.19	40.92	46.18	49.18
Multi-level	**55.50**	**67.51**	**78.26**	**82.75**	**85.00**	**36.15**	48.34	62.44	69.94	74.37

Table 3.5: The performance of using different levels of representation for few-shot learning on the same task (Generic object classification) and another different task (Fine-grained object classification). The top-5 accuracy of the novel categories in the novel label space (Novel/Novel) is reported. WC denotes our proposed weight-centric constraint. Best are bolded.

Method	n=1	2	5	10	20	Classifier
Baseline [116, 117]	53.96	62.88	69.55	71.56	73.42	81.80
Baseline + WC	**69.93**	**74.94**	**78.30**	**78.99**	**79.68**	81.71

Table 3.6: Top-1 Classification accuracy on CUB Base-class test set using samples as the classifier in two feature spaces.

Cross-domain performance of low-shot learning. We investigate the transferability of different levels of representations in the FSL setting. To achieve this, we perform a cross-domain evaluation, where we evaluate the learned model on both the same domain and different domain data. Specifically, we first train a model on the base-class data from the ImageNet dataset. Then we evaluated it on both ImageNet and the Caltech-UCSD bird dataset[24]. Table3.5 presents the comparison results obtained based on the Resnet-50 backbone and the Avg weight generator. First, we can observe that learning with weight-centric constraints improves performance in both the same-domain and cross-domain settings. Also, the mid-level features achieve the best accuracy in cross-domain testing, while the relation-level performs the worst. This result reveals that the mid-level representation exhibits strong transferability in the FSL setting. Furthermore, the proposed multi-level representation achieves the best accuracy on the same-domain data and obtains comparable performance with the mid-level features. This indicates that using multi-level features for FSL help improve generalization ability and handle domain shift problems.

3.4.4 Ablation study

Effiectiveness of the classifier-centric constraint. To verify the effectiveness of the classifier-centric constraint, we established the following experiments. First, we train two ConvNet models on the base class data, with and without classifier-centric constraints, to learn the two feature spaces. Then we randomly sample some samples from each class of the base class dataset to construct two classifiers to classify the test set. Finally, by evaluating their classification performance, it is indicated in which feature space the sample can construct a better decision boundary. The experimental results are shown in Table3.6. We can observe that the feature space learned with cosine softmax loss achieves poor accuracy which indicates the sample points in this space might be scattered and not close the classifier weight. By applying the classifier-centric constraint, the accuracy is significantly improved. This demonstrates that the feature space learned with classifier-centric constraints is more suitable for building classifiers using samples. We further evaluate the classifier-centric constraint under different evaluation metrics and provide the results in Fig3.6. We can see that our proposed constraint improves the

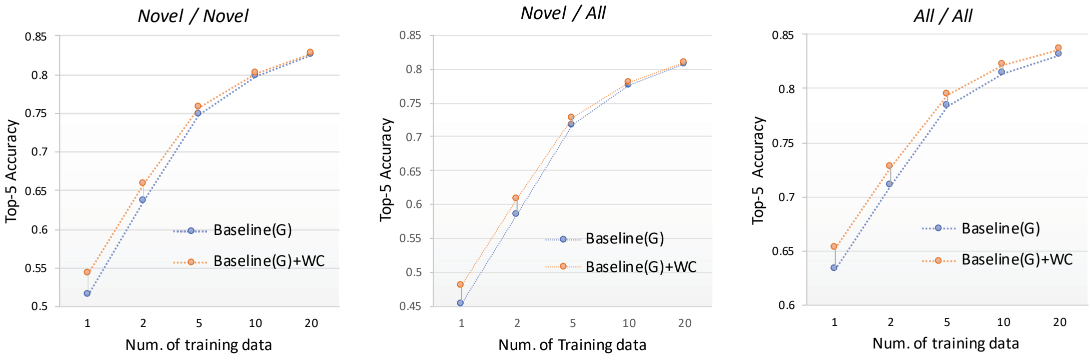


FIGURE 3.6: Top-1 Classification accuracy of few-shot setting on CUB set. Here, baseline refers to the feature space learned with cosine softmax loss, WC denotes our proposed weight-centric constrain.

	Novel / Novel			Novel / All		
	n=1	2	5	n=1	2	5
H(baseline)	51.56	63.67	74.78	45.26	58.53	71.80
H+WC	54.24	65.71	75.75	47.95	60.77	72.91
(H+WC)+M	56.96	68.50	78.58	50.79	64.30	76.52
(H+WC+M)+R	57.12	68.28	77.77	53.48	65.82	76.95

TABLE 3.7: Oblation study experiments on the ImageNet-based few-shot benchmark. *H*, *M*,and *R* refer to High-, Mid-, and relation-level features, respectively. *WC* refers to using a weight-centric learning strategy.

baseline consistently in three evaluation metrics. More importantly, the improvements under the "ALL/ALL" setting are the most significant, revealing that the classifier-centric constraint exhibits superiority in generalized few-shot learning. Fig3.7 shows a comparison of the intra-class variance between two feature spaces learned with and without weight-centric constraint. It can be seen that training with weight-constraint reduces intra-class variance on both training and test sets, and also both base and novel class data.

The contribution of each component. We conduct an ablation study to compare the performance of different levels of representation in the FSL setting. The table provides an ablation study on the Few-shot-imagenet benchmarks to observe the effect of each element. On the one hand, we can see that when evaluating only the novel label space, adding the weight-centric and mid-level components in sequence continuously improves the performance. This shows that both pieces help enhance the model generalization ability, which also implies that increasing the prototype-ability and transferability of feature representation can benefit few-shot learning. However, incorporating relational-level features does not further increase the performance in this setting. However, it shows a significant improvement under the "novel/all" evaluation metric. This indicates

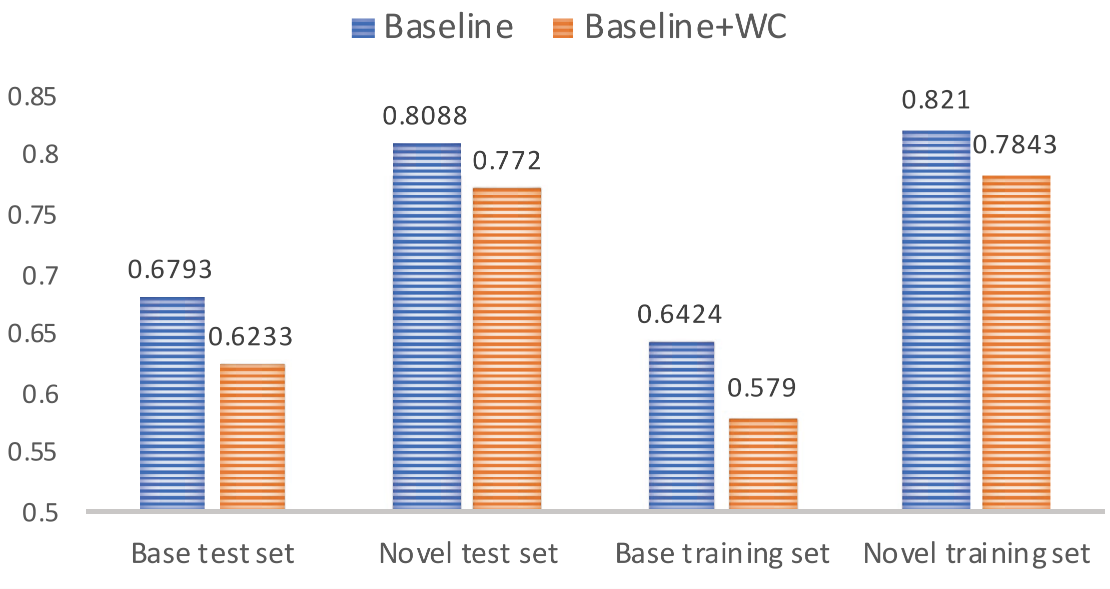


Figure 3.7: Comparison of the intra-class variance between two feature spaces both learned on the base training set. Here, baseline refers to the feature space learned with cosine softmax loss, and WC denotes our proposed weight-centric constrain. Note that we report the average intra-class variance for each dataset.

that the relation-level features have a weaker generalization to novel classes but can effectively prevent novel class data from being classified into the base categories.

We also provide some prediction results in Fig3.8, which can be used to intuitively analyze the few-shot learning ability of different representations. For example, the test images in the second column mostly contain some patterns(e.g. objects or parts of objects) which are very similar to those occur in the training examples, while the similarities between images in the last two columns and the training images tend to be subtle.

3.5 Conclusion

This work investigates the problem of feature representation in few-shot learning. To improve the representation power for unseen categories and weight generation capacity in feature learning, we proposed a multi-level weight-centric representation learning approach. The method first incorporated mid- and relation-level features with high-level to enhance representation capacity. In addition, a classifier-centric learning strategy was proposed to allow a few sample features to be used to construct a more discriminative classifier. Compared with existing methods, the method increases the feasibility of building a discriminative decision boundary based on a few samples. Also, it improves the transferability for characterizing novel classes and preserves the classification capability

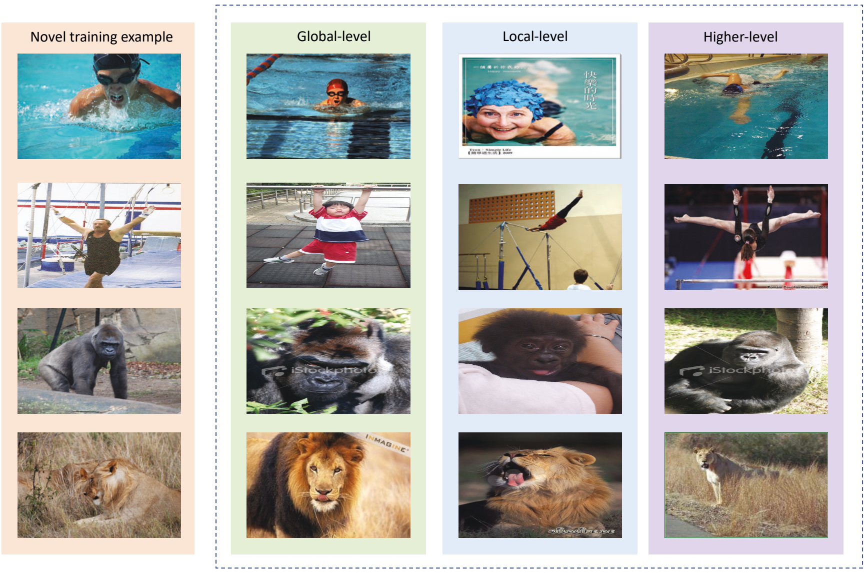


Figure 3.8: Some successful exemplars using our proposed method. The first column shows a single training image of a novel class, all images in the remaining three columns are correctly predicted by using the proposed multi-level representation. The second column shows some successful predictions using only global-level features but they are misclassified if using local or higher-level representation, and so on for the second and the third column.

for base classes. In experiments, we extensively evaluated our approach on two low-shot classification benchmarks and demonstrated its effectiveness in improving generalization. Our proposed method can also benefit other tasks, such as zero-shot learning and image retrieval, in which feature extractors play a critical role. However, one drawback of our approach is that it constructed multi-level features by concatenating multiple features, introducing redundancy in learning. Therefore, in future work, we will investigate how to learn a compact representation from numerous information sources. In addition, our proposed method may suffer from forgetting base-class knowledge when more novel classes are expanded into the classification model. Thus, our future work will investigate how to avoid forgetting issues in long-term incremental learning settings.

Chapter 4

Dynamic Semantic Structure Distillation for Low-resolution Fine-grained Recognition

4.1 Introduction

As introduced in Chapter 1, FGVC has made unprecedented progress with the rise of deep learning. In recent years, a variety of methodologies have emerged to address the challenges associated with fine-grained image classification, such as part-based strategies [14, 45, 46, 55] that meticulously extract details from images, and sampling-based methods[71, 72] that continuously enrich feature representations to improve the capacity to discern subtle differences between fine-grained categories.

Despite the impressive strides made in classification accuracy for fine-grained recognition tasks, these methods predominantly rely on high-resolution image inputs, such as those with dimensions of 448x448 pixels. Consequently, the performance of these models suffers dramatically when confronted with low-resolution image inputs. For instance, the MGE-CNN method has achieved an accuracy of 88.5% on the CUB dataset. However, when trained and tested on low-resolution image data, its accuracy drops to 62.95 %. This performance decline can be mainly attributed to two factors: 1). The significant loss of detailed information in low-resolution images; and 2). The pretraining model used by these methods lacks semantic prior in low-resolution images.

A straightforward solution to this problem is to use knowledge distillation methods to extract knowledge from high-resolution models to guide the learning of low-resolution models. Nevertheless, existing knowledge distillation approaches focus primarily on general category classification tasks, and the extracted knowledge is mainly related to single and static semantic structures. For example, the input images of teacher and student models mostly only contain objects of a single category. Therefore, it is not suitable for fine-grained image classification tasks. This is because different fine-grained objects usually share similar or identical parts, and semantic relations between parts of different categories are crucial for fine-grained category discrimination.

This chapter presents a novel learning framework called Dynamic Semantic Structure Distillation (DSSD) to address the above-mentioned issue. Our approach comprises two primary components: dynamic semantic structure learning and decoupled knowledge distillation. The former aims to train both teacher and student networks to perceive dynamic semantic structures effectively and can be implemented based on cutting and pasting data augmentation and a consistent semantic loss. The goal is to facilitate learning and distillation of dynamic semantic structure and part relations. The latter component focuses on the decoupling of network output into two distinct types of knowledge: semantic composition knowledge and non-target category distribution knowledge. These types of knowledge are then separately distilled from the teacher model to the student model. By doing so, our method is capable of effectively transferring the high-resolution model's semantic structure prior to the student model. Furthermore, this approach strengthens the student model's ability to distinguish semantic details of objects in low-resolution images.

We evaluated our proposed method on two distinct knowledge distillation tasks: high-to-low resolution and large-to-small model. Specifically, we first verify the effectiveness of our approach by comparing it experimentally with state-of-the-art fine-grained object recognition and knowledge distillation methods, using fine-grained classification datasets as benchmarks. The experimental results show that our method consistently outperforms existing methods on three datasets and four network backbones. In addition, we compare our approach with knowledge distillation methods on the task from large to small models and demonstrate its effectiveness on both general image classification and standard knowledge distillation tasks. Through our experiments, we validate the efficacy of our proposed method and demonstrate its superiority over existing techniques in terms of knowledge distillation performance. Overall, the results obtained from the evaluation of our method indicate its potential to significantly improve the performance of knowledge distillation tasks for both high-to-low resolution and large-to-small model scenarios. Our approach represents a promising avenue for future research in the field of image recognition and knowledge distillation.

4.2 Related work

4.2.1 Low-resolution task

Over the past few decades, the Low-resolution domain has gained substantial attention due to its challenges and broad applicability, particularly in Face recognition and Object detection. For Face recognition, Low-resolution (LR) techniques can be broadly classified into two groups: super-resolution[143–149] and feature-learning methods[150–152]. The first group mainly exploits a way to restore LR images to High-resolution(HR) images using computer vision synthesis algorithms. Although face hallucination[146] and simultaneous SR and recognition[144] have been proposed as typical works, most of these methods [147, 148, 153] focus on the visual fidelity of the generated images, but do not improve face recognition performance. In contrast, feature-learning methods directly extract discriminative features from LR images or transfer the feature space between HR and LR to address dimensional mismatch[151]. Recent years have seen the rapid development of LR for face recognition, with several works[145, 154, 155] emerging for the Very Low-resolution (VLR) task. For example, Yu et al. [155]presented a framework that utilizes facial attribute information in face super-resolution and achieved significant performance on the VLR task of 16×16 pixels. Singh et al. [156] explored a DirectCapsNet model to address the limited information contained in VLR images. In addition, many studies[157–160] have been devoted to developing the Low-resolution Object detection model. Lu Qi[158] combined an aligned multi-scale training and knowledge distillation to improve the performance for low-resolution instance-level detection.

4.2.2 Knowledge Distillation

Knowledge distillation(KD) is a model compression method based on the ”teacher-student-network” idea, which distills the knowledge from a large ”teacher” model into a small ”student” network. This theory was first proposed by Hinton et al. [161] in 2015 and is widely used in the industry for simplicity and effectiveness. There are different types of knowledge[162], for example, logits, feature, and relation-based. Logits-based methods[161, 163–165] learn the knowledge from the output layer of the ”teacher” model. In addition to the output layer,feature-based methods [166–169] transfer the feature representation of the middle layers as knowledge. Recent works[170–173]in the third group mainly explored the relationship between input-hidden-output layers and then ”teach” the output of the specific layers in the teacher network to the student model. However, most of these methods focus on optimizing compression models. They can not directly

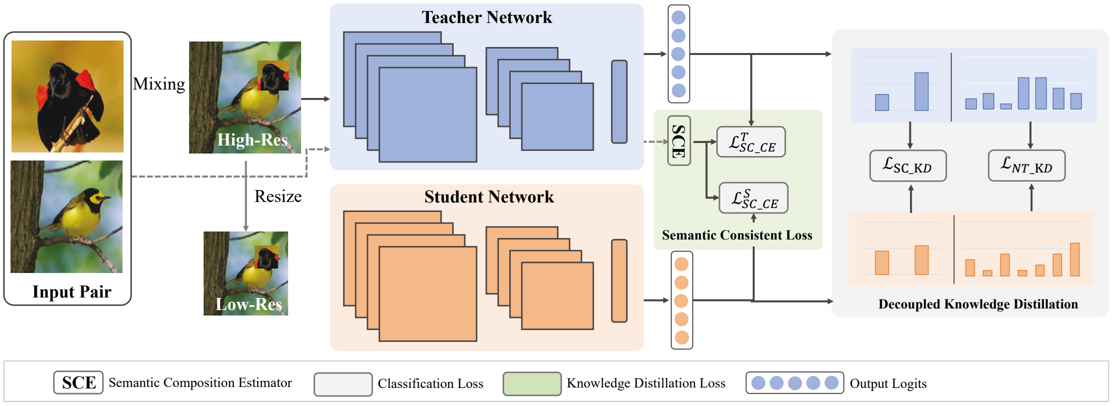


FIGURE 4.1: An overview of our proposed method ***Dynamic Semantic Structure Distillation (DSSD)***. Unlike existing knowledge distillation approaches, our method distills semantic structure knowledge and non-target category distribution separately by constructing the input with dynamic semantic structures and then decoupling the output.

apply to fine-grained recognition because of the characteristics of the subtle difference between fine-grained classes.

4.3 Methodology

In this chapter, we delve into our proposed method, Dynamic Semantic Structure Distillation (DSSD), illustrated in Fig.4.1. DSSD comprises two primary components: ***dynamic semantic structure learning*** and ***decoupled knowledge distillation***. The first component aims to enhance the understanding of dynamic semantic structures by the teacher and student networks. We achieve this through a data augmentation strategy that involves cut-and-paste operations while ensuring a consistent semantic loss. By leveraging this approach, both networks are encouraged to perceive and understand the changing semantic structures within the data. The second component, decoupled knowledge distillation, involves separating the network output into two distinct types of knowledge: semantic composition knowledge and non-target category distribution knowledge. This knowledge is then distilled from the teacher model to the student model in a differentiated manner. Specifically, the semantic composition knowledge is transferred to the student model, whereas the non-target category distribution knowledge is distilled separately.

4.3.1 Dynamic semantic structure learning

One limitation of conventional knowledge distillation methods is their reliance on static and single-semantic images and supervisory signals. Consequently, the knowledge learned and distilled by these methods strongly correlates with the semantic distribution of the training data, resulting in a weaker applicability to the test data. To address this limitation, here we encourage the network to learn and distill richer knowledge by utilizing mixing data augmentation strategies [174, 175] to construct data with dynamic and mixed semantic composition.

Input image preprocessing. Given a randomly chosen image pair (x_1, x_2), we generate the input $\tilde{x}$ using the asymmetric mixing operation [174], which is expressed as

$$\tilde{x} = (1 - M_{\lambda_1}) \odot x_1 + T(M_{\lambda_2} \odot x_2), \tag{4.1}$$

where M_{λ_1} and M_{λ_2} are binary masks processed based on the area ratios λ_1 and λ_2, and T is a transformation function and $\odot$ is element-wise multiplication.

Semantic composition estimator. To estimate the semantic composition of a mixed image, CutMix [175] calculates the area ratio of the regions involved in the mixing. However, the area ratio usually cannot reflect the semantic proportion in hybrid images. So, we follow the work [174] which uses the intensity of attention in the region to estimate the corresponding semantic ratio. Given an input image x, we first calculate the semantic ratio at pixel i location by:

$$SR(x^i) = \frac{CAM(x^i)}{\sum(CAM(x))}, \tag{4.2}$$

where $CAM(\cdot)$ is a function to obtain the class activation values. Then, we can estimate the semantic composition (s_1, s_2) for an mixed image $\tilde{x}$ using a semantic composition estimator ($\boldsymbol{SCE}$), which is defined as:

$$SCE(\tilde{x}) = (1 - \sum(M_{\lambda_1} \odot SR(x_1)), \sum(M_{\lambda_2} \odot SR(x_2))) \tag{4.3}$$

Dynamic semantic structure learning. The teacher model is usually pre-trained in a typical knowledge distillation learning framework, and the model weights are fixed during the knowledge distillation process. However, fixing the teacher model weights cannot acquire transitional knowledge during training. Therefore, we also train the teacher model from scratch during knowledge distillation. Here, the training loss for the teacher and student networks are

$$\ell^T_{SC\_CE} = s^T_1 CE(O^T(\tilde{x}^T), y_1) + s^T_2 CE(O^T(\tilde{x}^T), y_2), \tag{4.4}$$

$$\ell^S_{SC\_CE} = s^S_1 CE(O^S(\tilde{x}^S), y_1) + s^S_2 CE(O^S(\tilde{x}^S), y_2), \tag{4.5}$$

where $CE(\cdot)$ denotes cross-entropy loss, $O^T(\cdot)$ and $O^S(\cdot)$ are outputs of the teacher and student networks, respectively.

4.3.2 Decoupled Knowledge Distillation

Our method introduces an advanced framework that incorporates a decoupled knowledge distillation strategy that addresses the limitations of existing knowledge distillation loss functions. Traditional methods are adept at distilling static semantic knowledge, but fall short when it comes to dynamic and mixed semantic structures. This shortfall is mainly due to the significant weakening of compositional knowledge in mixed semantics caused by output softening. Our approach therefore implements a decoupled strategy in the knowledge distillation process.

In this strategy, we categorize network output into two distinct types:'semantic composition knowledge' and 'non-target category distribution knowledge.' This bifurcation is essential because it allows for a more precise and effective distillation process. The semantic composition knowledge, which forms the core understanding of the subject matter, is directly transferred to the student model. Meanwhile, the non-target category distribution knowledge, addressing broader context and supplementary information, is distilled separately.

For the technical implementation, we begin by decoupling the network output into primary semantic composition and non-target category distribution knowledge. Let z denote the network output logits and y_1 and y_2 the target labels for the two images used to create the blended image. We can then decouple the output z into semantic composition logits (z^{y_1}, z^{y_2}) and non-target category logits $z^{/y_1,y_2}$. These logits are transformed into probabilities (p^{y_1}, p^{y_2}) and $\mathbf{p}^{/y_1,y_2}$:

$$p^{y_1} = \frac{\exp(\frac{z^{y_1}}{\tau_{sc}})}{\exp(\frac{z^{y_1}}{\tau_{sc}}) + \exp(\frac{z^{y_2}}{\tau_{sc}})}, \tag{4.6}$$

$$p^{y_2} = \frac{\exp(\frac{z^{y_2}}{\tau_{sc}})}{\exp(\frac{z^{y_1}}{\tau_{sc}}) + \exp(\frac{z^{y_2}}{\tau_{sc}})}, \tag{4.7}$$

$$p_i^{/y_1,y_2} = \frac{\exp(\frac{z_i^{/y_1,y_2}}{\tau_{nt}})}{\sum_i(\exp(\frac{z_i^{/y_1,y_2}}{\tau_{nt}}))}, \tag{4.8}$$

where τ_{sc} and τ_{nt} are temperature parameters that modulate a softer probability distribution for semantic composition and non-target category logits, respectively.

To prevent the dilution of primary semantic knowledge due to the softening of the probability distribution, we transfer the semantic composition and non-target category knowledge separately from the teacher to the student model. The corresponding learning loss is calculated using the KL-divergence loss function:

$$\ell_{SC\_KD} = KL(p_T^{y_{1,2}} || p_S^{y_{1,2}}), \tag{4.9}$$

$$\ell_{NT\_KD} = KL(p_T^{/y_1,y_2} || p_S^{/y_1,y_2}), \tag{4.10}$$

where $KL(\cdot)$ represents the Kullback-Leibler divergence loss, measuring the similarity between two probability distributions.

Finally, the overall loss of our knowledge distillation framework is defined as:

$$\mathcal{L} = \ell_{SC\_CE}^S + \ell_{SC\_CE}^T + \eta\ell_{SC\_KD} + (1-\eta)\ell_{NT\_KD}, \tag{4.11}$$

where η is a hyperparameter that balances the importance between the semantic composition and non-target category distribution knowledge. This comprehensive loss function effectively ensures the nuanced transfer of both core and contextual knowledge to the student model, optimizing the distillation process for dynamic and mixed semantic structures. More details of the training process are summarized in **Algorithm** 2.

4.4 Experiments

In this chapter, we conduct a comprehensive investigation to understand the impact of primary hyperparameters on the performance of our model. Specifically, our experiments focus on the knowledge distillation task transitioning from high-resolution to low-resolution models. For this purpose, we used the Resnet50 network model and employed the CUB dataset for evaluation. In detailing our experimental setup, we specify the range and values of hyperparameters explored, the rationale behind their selection, and the method used for tuning them.

Algorithm 2: Dynamic Semantic Structure Distillation

```
Input: Training data $X = \{(x_i, y_i)|i \in [0, 1, ..., N-1]\}$
Hyperparameters $[\beta, \tau_{sc}, \tau_{nt}, \eta, ]$
Parameter: Network parameters $[\Theta_T, \Theta_S]$
Output: Trained parameters $\tilde{\Theta}_S$
Network initialization $Net_{\Theta_T}$, $Net_{\Theta_S}$.
while Not Converge do
    Sample pair data $((x_1, y_1), (x_2, y_2)) \leftarrow X$
    Sample area ratio $(\lambda_1, \lambda_2) \leftarrow Beta(\beta, \beta)$
    $\tilde{x} \leftarrow Gen(x_1, x_2, \lambda_1, \lambda_2)$
    $s_1^T, s_2^T \leftarrow SCE(x_1, x_2, \lambda_1, \lambda_2, Net_{\Theta_T})$
    $s_1^S, s_2^S \leftarrow SCE(x_1, x_2, \lambda_1, \lambda_2, Net_{\Theta_S})$
    $z^T \leftarrow Net_{\Theta_T}(\tilde{x})$
    $z^S \leftarrow Net_{\Theta_S}(\tilde{x})$
    Compute loss $\mathcal{L}$ from Eq.4.11
    $update(\Theta_T, \Theta_S) \leftarrow \frac{\partial \mathcal{L}}{\Theta_T}, \frac{\partial \mathcal{L}}{\Theta_S}$
end while
return Trained parameters $\tilde{\Theta}_S$
```

Subsequently, we delve into a component-wise analysis of our proposed method. This involves systematically evaluating each component's contribution to performance enhancement in knowledge distillation. We describe the experimental design for isolating the effects of individual components, detailing the configuration variations, and the metrics used to assess their impact.

Finally, we demonstrate the effectiveness of our method in two key scenarios: knowledge distillation from high- to low-resolution models and from large- to small-models. In each scenario, we provide a detailed account of the experimental conditions, including the models compared, the datasets used, and the specific challenges these scenarios present. Furthermore, we present a thorough analysis of the results, highlighting key findings and insights that showcase our method's strengths and potential areas for improvement.

Throughout our experimental analysis, we aim to provide a clear and comprehensive understanding of the performance of our method under various conditions. This includes discussing any surprising or counterintuitive findings and how they contribute to our understanding of knowledge distillation in different resolution and model size contexts. In doing so, we ensure that our experimental study not only validates the effectiveness of our approach, but also offers valuable insights into its underlying mechanisms and potential applications.

High-to-low resolution: For this task, a teacher model is trained on high-resolution data and knowledge is distilled to the student model on low-resolution input. We specify employ an image resolution of 448×448 for the teacher model and 128×128 for the student model. The teacher and student models share the same network structure.

In addition, we utilize a variety of network structures (including Resnet 50,34,50 and 101)to validate our method's applicability in terms of network depth. The effectiveness of our proposed method is demonstrated through an experimental comparison with current state-of-the-art fine-grained image classification techniques, as well as knowledge distillation methods applied to fine-grained image classification datasets.

Large-to-small model: This task aims to transfer knowledge from a large network model to a small deep learning model. We follow the work [176] and use ResNet56, ResNet110, ResNet32×4, WRN-40-2, WRN-40-2, VGG13 as model teachers. Correspondingly, we utilize a set of student models, specifically ResNet20 ResNet32, ResNet8×4, WRN-16 -2, WRN-40-1 VGG8. Experimental evaluations are conducted on a standard image classification data set, in which the performance of our approach is compared with that of contemporary state-of-the-art methods.

4.4.1 Training Details

We implemented our method based on PyTorch and trained it on a Nvidia V100 GPU. In training, we initialized backbone networks with Imagenet pre-trained models for fine-grained recognition tasks. We used stochastic gradient descent (SGD) with a momentum of 0.9, the base learning rate of 0.001 for the pre-trained weights, and 0.01 for new parameters. We trained our model for 200 epochs and decayed the learning rate by a factor of 0.1 every 80 epochs. We employed some common practices for fine-grained datasets in training. We adopted standard data augmentations, including random cropping and random horizontal flipping.

4.4.2 Results and analysis

Influence of hyperparameters β and η. The hyperparameter β determines a Beta distribution used to sample a random patch in the generation of images with mixed semantics. We evaluated its impact by testing the values β [list specific values]. As shown in Fig.4.2, the performance of the model gradually improved with increasing β values, reaching a peak at 3. This indicates the significance of medium-size patch mixing in the creation of effective mixed semantic images. Interestingly, overall accuracy remains relatively stable across different β values, suggesting the resilience of our framework to variations in β. For η, controlling the balance between mixed semantic loss and non-target distribution distillation loss, Fig. 4.3 reveals that higher η values, favoring mixed semantic loss, lead to reduced distillation effectiveness. Conversely, increasing the weight of non-target distribution loss enhances performance, underlining the importance of non-target distribution knowledge in our distillation process.

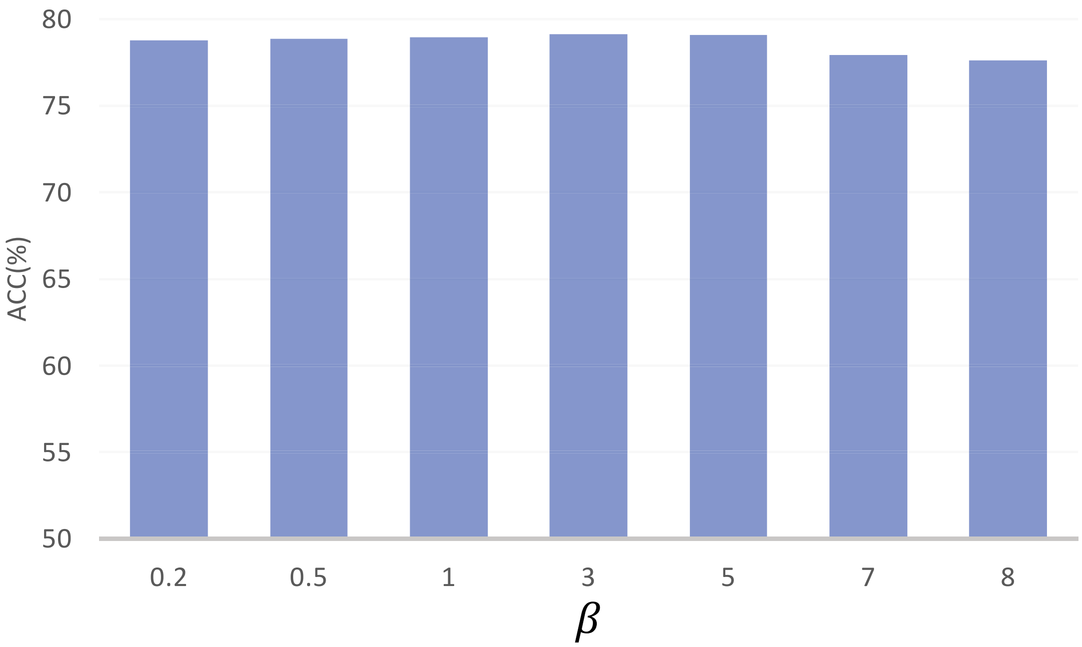


FIGURE 4.2: Ablation study of the hyperparameters β for dynamic semantic structure learning. The results illustrate that our method is not sensitive to β and performs consistently well across different β values.

	CUB		Aircraft	
	Res-50	Res-101	Res-50	Res-101
Baseline	65.87	68.12	57.96	65.14
+DSSL	71.32	73.04	63.87	69.29
+KD	71.14	72.71	58.22	64.68
+DSSL+KD	72.35	74.43	64.31	70.35
+DSSL+DKD	**79.13**	**81.07**	**73.01**	**78.99**

TABLE 4.1: Ablation study (Acc.%) of each component. DSSL and DKD are the abbreviations of Dynamic Semantic Structure Learning and Decoupled Knowledge Distillation, respectively, and KD is the abbreviation of the classic algorithm Knowledge Distillation.

Effectiveness of each component of DSSD. In assessing DSSD's components, DSSL (Dynamic Semantic Structure Learning) and DKD (Decoupled Knowledge Distillation), we first established a baseline by training a standard network on low-resolution data. We then constructed high-resolution teacher and low-resolution student models using the same network and applied classic knowledge distillation (KD) for benchmarking. Subsequently, we evaluated the impact of integrating DSSL and DKD individually and in combination. Table 4.1 shows that both DSSL and KD independently enhance baseline performance. However, the synergy of DSSL with DKD yields a more significant performance boost than KD alone, highlighting DKD's superior efficacy in distilling dynamic mixed semantic knowledge.

Knowledge distillation from high to low-resolution model. In Tables 4.2 and 4.3,

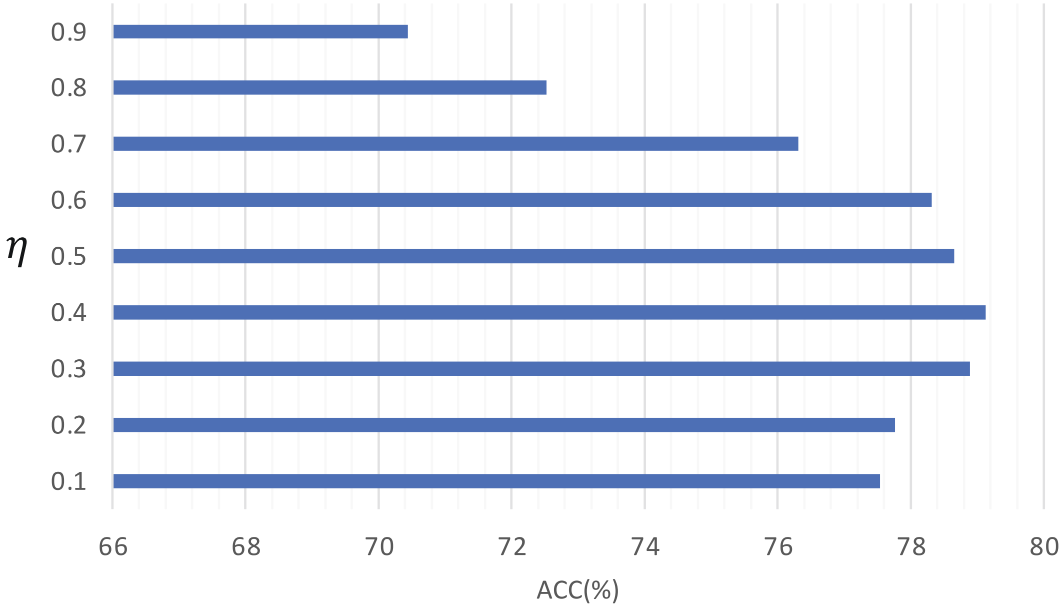


Figure 4.3: Ablation study of the hyperparameters η for decoupled knowledge loss.

	CUB		Aircraft		Cars	
	ResNet-18	ResNet-34	ResNet-18	ResNet-34	ResNet-18	ResNet-34
Baseline	58.98	63.01	50.67	51.83	61.68	63.78
SnapMix [174]	64.98	67.13	58.67	61.79	72.14	76.84
KD [161]	63.15	64.88	53.47	56.04	60.58	64.83
RKD [173]	61.84	64.78	53.35	55.18	60.87	64.38
CRD [177]	64.11	66.12	54.31	56.23	61.72	65.79
OFD [178]	63.91	65.83	54.69	55.67	61.37	65.63
ReviewKD [179]	67.89	67.62	58.54	58.79	71.02	73.12
DKD [176]	70.97	68.13	58.04	60.12	75.43	78.96
DSSD	**74.13**	**70.21**	**66.32**	**70.01**	**81.52**	**86.24**

Table 4.2: Performance comparison with state-of-the-art knowledge distillation methods(Acc.%). This table shows the results of distilling knowledge from the high-resolution teacher model to the low-resolution student model. Two backbone networks, Resnet-18 and Resnet-34, and three fine-grained datasets are used for each method for a comprehensive comparison.

we compare our method with state-of-the-art knowledge distillation techniques in the transition from high-resolution to low-resolution models. We re-trained these methods for this specific task, as they were originally reported for large-to-small-model tasks. The results in Table 4.2 show significant performance degradation when directly training and testing shallow networks on low-resolution datasets. Our method outperforms others in various settings in the CUB, aircraft, and car datasets, as detailed in Table 4.3. For deep networks like Resnet101, our approach shows improvements of 5.4%, 11.2%, and 3.3% respectively, compared to the latest method DKD. Furthermore, as shown in Table 4.4,

	CUB		Aircraft		Cars	
	ResNet-50	ResNet-101	ResNet-50	ResNet-101	ResNet-50	ResNet-101
Baseline	65.87	68.12	57.96	65.14	68.85	71.84
SnapMix [174]	70.78	72.59	64.07	68.89	82.13	83.96
KD [161]	70.84	71.98	58.76	66.11	71.83	76.12
RKD [173]	70.09	71.33	57.63	64.67	71.87	74.63
CRD [177]	72.83	72.55	58.77	66.39	72.02	75.32
OFD [178]	72.71	71.97	58.63	65.67	71.97	74.83
ReviewKD [179]	73.81	75.52	67.65	68.89	84.82	85.72
DKD [176]	76.02	75.68	61.89	68.54	85.04	84.89
DSSD	**79.13**	**81.07**	**73.01**	**78.99**	**88.12**	**89.13**

Table 4.3: Performance comparison with state-of-the-art knowledge distillation methods(Acc.%). This table shows the results of distilling knowledge from the high-resolution teacher model to the low-resolution student model. Two deeper backbone networks, Resnet-50 and Resnet-101, and three fine-grained datasets are used for each method for a comprehensive comparison.

Method	Accuracy(%)		
	CUB	Cars	Aircraft
PCA-Net [180]	64.87	68.16	79.92
PMG [16]	61.01	65.13	61.87
API-Net [96]	69.05	73.97	64.51
MGN-CNN [15]	62.95	70.59	68.21
ResNet-50	65.84	69.13	59.24
ResNet-101	68.25	71.85	65.28
Res-50+DSSD	78.29(+12.45)	87.32(+18.19)	71.89(+12.65)
Res-101+DSSD	**80.29(+12.04)**	**87.96(+16.11)**	**79.08(+13.80)**

Table 4.4: Performance comparison with fine-grained methods(Acc.%). All comparing approaches are tested on low-resolution images (128x128). The numbers in brackets indicate the value of performance improvement of our approach over the baseline model.

our method significantly exceeds those tailored for fine-grained image classification in low-resolution image contexts.

Knowledge distillation from large to small model. Focusing on the large-to-small model task, we evaluated existing knowledge distillation methods using the CIFAR100 dataset, as detailed in Table 4.5. Our method consistently achieved superior performance under six different model configurations. The performance improvement observed in this task is somewhat less pronounced than in the high- to low-resolution tasks. This may be attributed to the reliance of CIFAR100 on global features for category distinction, which poses a lower demand for combinatorial knowledge dependencies compared to fine-grained classification tasks.

Visualization. To gain intuition about the superiority of our method, we visualize and compare the attention maps of the low-resolution models. As can be seen from Figure 4.4, training directly from low-resolution data, the learned model (baseline) pays

Method		Res56	Res110	Res32 × 4	WRN-40-2	WRN-40-2	VGG13
	Teacher	72.34	74.31	79.42	75.61	75.61	74.64
		Res20	Res 32	Res8 × 4	WRN-16-2	WRN40-1	VGG8
	Student	69.09	71.14	72.50	73.26	71.98	70.36
Features	FitNet [166]	69.21	71.06	73.50	73.58	72.24	71.02
	RKD [173]	69.61	71.82	71.90	73.35	72.22	71.48
	CRD [177]	71.16	73.48	75.51	75.48	74.14	73.94
	OFD [178]	70.98	73.23	74.95	75.24	74.33	73.95
	ReviewKD [179]	71.98	73.89	75.63	76.12	75.09	74.84
Logits	KD [161]	70.66	73.08	73.33	74.92	73.54	72.98
	DKD [176]	71.97	74.11	76.32	76.24	74.81	74.68
Logits	**DSSD(ours)**	**73.85**	**75.64**	**77.28**	**77.39**	**76.03**	**75.07**

Table 4.5: Performance comparison with state-of-the-art knowledge distillation methods on CIFA100 dataset(Acc.%). This table shows the results of distilling knowledge from the large teacher model to the small student model.

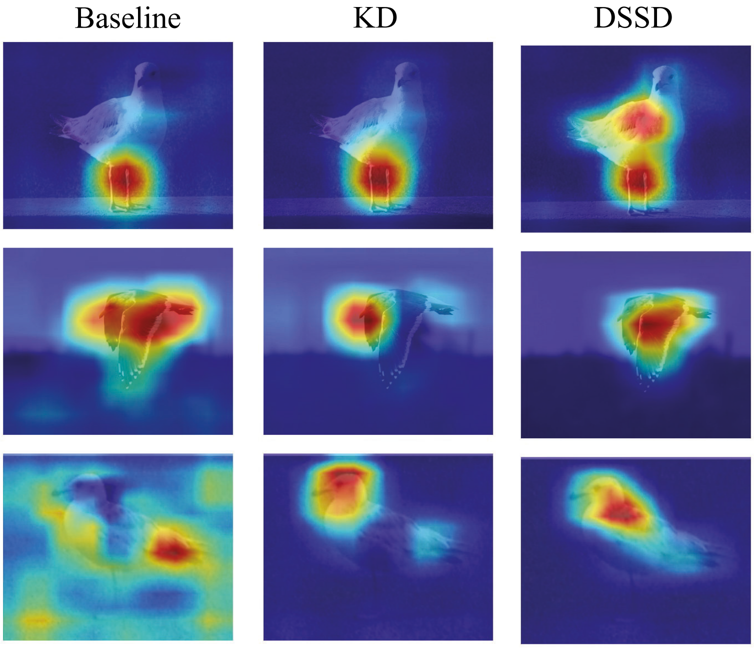


Figure 4.4: Attention map visualization of the distilled student model. The above attention maps are obtained by inputting low-resolution images. Among them, Baseline means that the model is directly trained and tested on low-resolution images, KD represents the results of the student model after using the classic knowledge distillation method KD, and DSSD is the visualization result of the student model learned by our method.

less attention to the main semantic regions, and the attention is easily distracted to the background regions. When the KD algorithm is used to extract the knowledge of a high-resolution model from low-resolution models, the attention distraction problem can be effectively solved. However, attention to the main semantic regions is still missing. After applying our method, the attention of the student model can better cover the main semantic regions.

4.5 Conclusion

In this chapter, we focus on the problem of low-resolution fine-grained recognition. We present a Dynamic Semantic Structure Distillation (DSSD) learning framework to address this issue. The method consists mainly of two main components: dynamic semantic structure learning and decoupled knowledge distillation. The former aims to train teacher and student networks to perceive dynamic semantic structures, while the latter decouples the network output into semantic composition knowledge and non-target category distribution knowledge to ensure effective knowledge distillation. First, we investigate the influence of primary hyperparameters and the effectiveness of each component in our method. Then, we demonstrate that our method consistently outperforms existing methods on three datasets and four network backbones. In addition, our approach is also practical in general image classification and standard knowledge distillation tasks.

Limitation and future work. One of the important limitations of our approach is that both the teacher and student models must be trained simultaneously, thus requiring a longer training time and computing resources. In the future, we will explore how to effectively distill semantic component relation knowledge to the student model under the premise of fixing the teacher model.

Chapter 5

Discussion

This chapter offers a critical analysis of the thesis's technical contributions, positioning them within the broader landscape of fine-grained image classification (FGVC) research. We analyze the methodological implications, strengths and limitations, and connect our findings to relevant literature. Furthermore, we propose directions for future work to extend and refine the presented approaches.

5.1 Strengths and Contributions

5.1.1 Comprehensive Survey as a Foundation

Chapter 2 of this thesis presents a structured literature review that categorizes FGVC methods into seven paradigms, informed by methodological trends and chronological development. This differs from conventional algorithmic taxonomies (e.g., [181]) by emphasizing the evolution of ideas from 2010 to 2023. For instance, while transformer-based models often overlap with enhanced representation learning, our categorization prioritizes their primary innovation trajectory. This survey provided both historical grounding and theoretical motivation for the thesis' methodological developments.

5.1.2 Advancing Few-Shot Learning

The multi-level weight-centric representation learning approach advanced the field by enabling **discriminative feature extraction and classification with limited samples**. The incorporation of mid-level and relation-level features significantly improved representation capacity, while the classifier-centric strategy improved decision boundary construction. Our method demonstrated **superior generalization** on low-shot classification benchmarks and proved its adaptability to tasks like **zero-shot learning** and **image retrieval**, showcasing its versatility.

Unlike prior works focused on single-level representations [115, 132], our method incorporates mid-level and relation-level semantics, promoting generalization with minimal supervision. Additionally, the classifier-centric strategy refines decision boundaries, a concept rarely explored in earlier few-shot frameworks.

5.1.3 Improving Low-Resolution Fine-Grained Recognition

The **Dynamic Semantic Structure Distillation (DSSD) framework** effectively addressed the challenge of low-resolution image classification. By integrating **dynamic semantic structure learning** and **decoupled knowledge distillation**, our approach outperformed state-of-the-art methods across various datasets and network backbones. This highlights its **scalability and practical applicability**, particularly in real-world scenarios where high-resolution data may not be readily available.

5.2 Limitations and Challenges

While the proposed methods demonstrate promising performance in addressing data inefficiency in FGVC, several limitations constrain their applicability and generalization. One significant limitation arises from the feature concatenation strategy employed in the weight-centric approach. Although this technique enhances representational richness by integrating multiple levels of features, it inadvertently introduces redundancy that may degrade learning efficiency and inflate model size. This issue becomes especially salient in high-dimensional feature spaces or when deploying models on edge devices where computational resources are constrained.

Additionally, the Dynamic Semantic Structure Distillation (DSSD) framework, despite its demonstrated effectiveness, incurs a high computational cost due to the requirement of simultaneous training for both teacher and student networks. This limitation poses challenges for scalability and real-time inference, making it less suitable for applications requiring fast deployment or adaptation on-the-fly.

A broader challenge shared by both methods is the limited capacity to generalize across domains. Specifically, the models exhibit performance degradation when exposed to unseen domains characterized by differences in lighting conditions, background clutter, or object scale. This suggests an over-reliance on the distributional characteristics of the training data and underscores the necessity for more domain-invariant or adaptive learning strategies. These challenges, while not unique to the proposed methods, reflect common issues in FGVC and highlight directions for refinement.

5.3 Implications for FGVC Research and Applications

The contributions of this thesis carry several broader implications for both the research community and the real-world deployment of FGVC systems. First, the effectiveness of multi-level feature representations and dynamic semantic structure modeling underscores the importance of hierarchical and context-aware representations in fine-grained recognition. These findings align with recent trends in computer vision literature suggesting that capturing inter-part relationships and semantic hierarchies is crucial for distinguishing subtle visual differences between closely related classes.

Moreover, by addressing two seemingly distinct challenges—few-shot learning and low-resolution classification—within a unified framework, this thesis highlights the potential for integrative approaches that simultaneously optimize for data efficiency and visual fidelity. This convergence opens up new opportunities for research aimed at solving multiple constraints jointly, for instance, through multitasking or modular learning paradigms.

In terms of practical impact, the emphasis on robustness, interpretability, and computational efficiency aligns well with application domains such as biodiversity monitoring, medical diagnostics, and industrial inspection. In these scenarios, high intra-class similarity, limited training data, and hardware constraints are common. Thus, this work not only advances academic understanding but also contributes insights that are directly translatable to field-deployable systems. Importantly, the observed limitations also serve as a reminder of the gap between benchmark performance and real-world robustness, calling for continued efforts in validation under diverse operational conditions.

5.4 Open Questions and Future Directions

The limitations and challenges identified in this thesis open several avenues for future research:

- **How can compact feature representations be constructed to eliminate redundancy while preserving discriminative power?** One major issue in feature learning is **redundant feature concatenation**, which increases computational complexity without necessarily improving discriminative power. Future research should explore:

  - **Learnable compression mechanisms**, such as **autoencoders or knowledge distillation**, to extract the most salient features while reducing redundancy.
  - **Sparse feature selection methods**, where only the most informative features contribute to classification, improving both efficiency and interpretability.
  - **Graph-based feature aggregation**, where relationships between feature components are explicitly modeled to refine representations.

- **How can efficient knowledge distillation techniques be designed to reduce computational overhead while maintaining performance?** Knowledge distillation is a powerful tool, but traditional methods require **expensive training of large teacher-student models simultaneously**. To improve efficiency, future work can:

  - Explore **one-shot or incremental distillation**, where knowledge is transferred gradually rather than requiring full teacher-student training.
  - Develop **adaptive distillation strategies**, where the student selectively learns important patterns rather than blindly mimicking the teacher's entire feature space.
  - Investigate **self-distillation** methods, where a single model refines its own predictions through progressive learning without external teacher supervision.

- **What strategies can enhance FGVC model robustness to variations in image quality, such as resolution, occlusion, and noise?** Fine-grained recognition models struggle with real-world conditions where images may be **low-quality, occluded, or contain background noise**. Potential solutions include:

  - **Adaptive attention mechanisms** that focus on the most informative regions of an image regardless of **resolution degradation or occlusions**.
  - **Contrastive learning for robustness**, where models learn invariant representations across different augmentations of the same object category.
  - **Domain adaptation techniques** that enable models to generalize better to new datasets with different lighting, textures, or backgrounds.

- **How can interpretability in FGVC models be improved to provide clearer insights into decision-making processes?** Interpretability is a growing concern in deep learning, especially in **critical applications like medical diagnostics and autonomous systems**. Future research should explore:
  - **Prototype-based learning**, where models learn from a small set of representative examples and make predictions based on similarity to known prototypes.
  - **Visualization techniques**, such as attention heatmaps, to provide human-interpretable justifications for model decisions.
  - **Hybrid approaches combining rule-based logic with deep learning**, allowing users to trace back decisions and understand the reasoning process in a structured manner.

By addressing these open questions, future research can build upon the contributions of this thesis to further advance FGVC. The proposed methodologies and insights provide a strong foundation for developing more **efficient, interpretable, and robust fine-grained classification systems**, paving the way for **real-world deployment in diverse application domains**.

Chapter 6

Conclusion and Future Work

This chapter presents a summary of the proposed methods and highlights their key contributions, performance, and limitations. Following this, we discuss potential directions for future research, focusing on both algorithmic enhancements and broader application possibilities.

6.1 Conclusion

This thesis addresses the overarching challenge of data inefficiency in fine-grained object classification (FGVC) through a progressive exploration that builds on a foundation established by a comprehensive survey of the field. Our contributions span three key aspects: an in-depth literature review that identifies fundamental challenges in FGVC, the development of a novel few-shot learning approach, and the proposal of an efficient knowledge distillation framework for low-resolution fine-grained recognition.

The survey provided a structured review of seven key paradigms in FGVC: Human in the Loop, Part Alignment-Based, High-Order Feature Learning, Multiple Granularity Features, Data Augmentation, Rich Representation Learning, and Transformer-Based Approaches. By analyzing these approaches, we identified critical gaps in the field, including the need for high-quality datasets, improved model generalization to unseen data, and robustness to image quality variations. This survey not only highlighted the limitations of existing methods but also laid the groundwork for our proposed solutions.

Building on these insights, we first tackled the issue of few-shot learning in FGVC by proposing a multi-level weight-centric representation learning approach. This method enhances representation capacity by integrating mid- and relation-level features and introduces a classifier-centric training strategy to construct more discriminative decision boundaries with limited data. Extensive experiments validated its effectiveness, demonstrating improvements in generalization for novel categories and applicability to tasks such as zero-shot learning and image retrieval. Despite these advancements, challenges such as feature redundancy, scalability in real-world scenarios, and knowledge retention in incremental learning remain open problems that require further investigation.

Next, we addressed the problem of low-resolution fine-grained image classification by developing a Dynamic Semantic Structure Distillation (DSSD) framework. This approach integrates dynamic semantic structure learning with decoupled knowledge distillation, allowing teacher and student networks to effectively capture fine-grained semantic information while reducing reliance on high-resolution images. Our experiments demonstrated consistent performance gains across multiple datasets and backbones, reinforcing its adaptability to both general image classification and standard knowledge distillation tasks. However, the requirement of simultaneous teacher-student training introduces computational overhead, which limits its feasibility for resource-constrained environments. Future work will focus on developing more efficient distillation strategies that enable knowledge transfer without requiring real-time co-training with large teacher models.

Together, this thesis makes significant contributions to advancing FGVC by systematically identifying and addressing key challenges in data efficiency and model scalability. The proposed methodologies bridge gaps in few-shot learning and low-resolution classification, providing new perspectives on multi-level representation learning, structured knowledge transfer, and efficient model adaptation. These contributions lay the groundwork for further research aimed at improving the interpretability, robustness, and scalability of fine-grained classification systems.

6.2 Future Work

While the methods proposed in this thesis advance FGVC research in several key areas, several open challenges remain that warrant further exploration. Future research directions can be categorized into three main areas: algorithmic enhancements, robustness and generalization, and practical deployment and scalability.

6.2.1 Algorithmic Enhancements

- **Feature Compactness and Discriminability:** The weight-centric learning approach demonstrated strong performance in few-shot settings, but feature concatenation introduces redundancy. Future work can explore feature selection and compression mechanisms, such as low-rank representations or attention-based feature refinement, to eliminate redundancy while preserving discriminative power.

- **Efficient Knowledge Distillation:** The DSSD framework improves low-resolution recognition but requires computationally intensive joint training of teacher and student models. A promising direction is one-shot or incremental distillation, where the student model progressively learns from a fixed teacher without requiring retraining.

- **Adaptive Model Learning:** Both proposed methods rely on static learning pipelines, but real-world applications often require continuous adaptation. Future research can develop self-supervised learning strategies that allow models to refine their representations dynamically without the need for extensive labeled data.

6.2.2 Robustness and Generalization

- **Domain Adaptation for Unseen Categories:** Generalization remains a challenge, particularly when dealing with domain shifts such as variations in lighting, occlusions, and background noise. Future research can explore contrastive learning, domain adaptation techniques, or meta-learning strategies to improve cross-domain generalization.

- **Enhancing Model Interpretability:** Fine-grained classification models often operate as black boxes, making it difficult to understand their decision-making process. Future work should integrate explainable AI (XAI) techniques, such as attention-based visualizations, prototype-based learning, or class activation mappings, to enhance model transparency and trustworthiness.

- **Robustness to Noisy Labels and Unreliable Annotations:** Many FGVC datasets suffer from label noise or inconsistencies in annotation. Future models should incorporate uncertainty estimation and self-correcting mechanisms to mitigate the impact of noisy labels.

6.2.3 Practical Deployment and Scalability

- **Lightweight FGVC Models for Real-Time Applications:** Many existing fine-grained classification models, including those proposed in this thesis, rely on deep architectures that may not be suitable for real-time or mobile applications. Future work should focus on model compression techniques, such as quantization, pruning, and lightweight neural network architectures, to enable deployment in resource-constrained environments such as embedded systems and edge devices.

- **Human-in-the-Loop FGVC Systems:** Despite advancements in automated classification, human expertise remains crucial in many fine-grained recognition applications (e.g., medical diagnosis, species identification, and industrial quality control). Future research could explore interactive learning frameworks where human experts guide model refinement through feedback loops.

- **Scalability to Large-Scale FGVC Datasets:** While the proposed methods perform well on benchmark datasets, scalability to large-scale, real-world fine-grained datasets remains an open challenge. Efficient training paradigms, such as distributed learning and self-supervised pretraining on large-scale unlabeled datasets, should be explored.

6.3 Final Remarks

In the broader landscape of computer vision, the research presented in this thesis contributes to a growing body of work seeking to improve data efficiency and adaptability in recognition tasks. While our methods target the specific challenges of few-shot and low-resolution fine-grained classification, their design principles—such as multi-level representation learning and semantic structure distillation—are applicable to other visual domains. These approaches align with a broader trend in computer vision towards learning transferable and structured representations that can generalize across tasks and domains.

Importantly, the rapid development of large-scale pretrained vision models, such as CLIP, DINO, and SAM, has redefined the baseline for many visual tasks. These models demonstrate remarkable zero-shot and cross-domain capabilities, primarily through self-supervised or vision-language pretraining. Our work complements this trajectory by addressing scenarios where pretrained models alone may fall short, such as in low-resolution inputs or highly imbalanced fine-grained categories where general-purpose representations are insufficiently specific. Future work may explore integrating our techniques—particularly the semantic-aware distillation and classifier-centric few-shot learning—with pretrained backbones to further enhance performance in limited-data regimes.

In this context, the methodologies proposed in this thesis can be seen not only as standalone solutions but also as complementary components that can enhance or refine the capabilities of general-purpose vision models. As pretrained models become more ubiquitous, the need for modular, task-adaptive, and efficient fine-tuning strategies like those proposed here will only grow.

Bibliography

[1] Yuning Chai. *Advances in fine-grained visual categorization*. PhD thesis, Oxford University, UK, 2015.

[2] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. Rpc: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019.

[3] Ibm ai algorithms can read chest x-rays at resident radiologist levels. https://www.ibm.com/blogs/research/2020/11/ai-x-rays-for-radiologists. Accessed: 2021-10-02.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

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

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[9] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, pages 438–451. Springer, 2010.

[10] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *2011 International Conference on Computer Vision*, pages 161–168. IEEE, 2011.

[11] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1173–1182, 2016.

[12] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014.

[13] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 365–374, 2017.

[14] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[15] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8331–8340, 2019.

[16] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020.

[17] Maria-Elena Nilsback. *An automatic visual flora-segmentation and classification of flower images*. PhD thesis, Oxford University, 2009.

[18] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[19] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 620–629, 2021.

[20] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[22] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[23] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

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

[25] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[26] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

[28] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[29] Michael Sperazza, Johnnie N Moore, and Marc S Hendrix. High-resolution particle size analysis of naturally occurring very fine-grained sediment through laser diffractometry. *Journal of Sedimentary Research*, 74(5):736–743, 2004.

[30] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for

generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[31] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1666–1674, 2015.

[32] Kevin J Shih, Arun Mallya, Saurabh Singh, and Derek Hoiem. Part localization using multi-proposal consensus for fine-grained categorization. *arXiv preprint arXiv:1507.06332*, 2015.

[33] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[34] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4148–4157, 2018.

[35] Kunran Xu, Rui Lai, Lin Gu, and Yishi Li. Multiresolution discriminative mixup network for fine-grained visual categorization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[36] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[37] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531. IEEE, 2011.

[38] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2013.

[39] Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, and Serge Belongie. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2014.

[40] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in

the loop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1153–1162, 2016.

[41] Ning Zhang, Ryan Farrell, and Trever Darrell. Pose pooling kernels for subcategory recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3665–3672. IEEE, 2012.

[42] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–736, 2013.

[43] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5546–5555, 2015.

[44] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1143–1151, 2015.

[45] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1134–1142, 2016.

[46] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

[47] Xuhui Yang, Yaowei Wang, Ke Chen, et al. Fine-grained object classification via self-supervised pose alignment. In *Computer Vision and Pattern Recognition*, 2022.

[48] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[49] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[50] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 511–520, 2017.

[51] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 2399–2406, 2015.

[52] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Augmenting strong supervision using web data for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 2524–2532, 2015.

[53] Lianbo Zhang, Shaoli Huang, and Wei Liu. Intra-class part swapping for fine-grained image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3209–3218, 2021.

[54] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1628–1636, 2021.

[55] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.

[56] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018.

[57] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8242–8251, 2019.

[58] Zhenchao Tang, Hualin Yang, and Calvin Yu-Chian Chen. Weakly supervised posture mining for fine-grained classification. In *Computer Vision and Pattern Recognition*, 2023.

[59] Ju He, Jie-Neng Chen, Shuai Liu, et al. Transfg: A transformer architecture for fine-grained recognition. In *AAAI Conference on Artificial Intelligence*, 2022.

[60] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv:2107.02341*, 2021.

[61] Yunqing Hu, Xuan Jin, Yin Zhang, et al. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *International Conference on Multimedia*, 2021.

[62] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. Vit-net: Interpretable vision transformers with neural tree decoder. In *International Conference on Machine Learning*, 2022.

[63] Qin Xu, Jiahui Wang, Bo Jiang, et al. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 2023.

[64] Peter N Belhumeur, Daozheng Chen, Steven Feiner, David W Jacobs, W John Kress, Haibin Ling, Ida Lopez, Ravi Ramamoorthi, Sameer Sheorey, Sean White, et al. Searching the world's herbaria: A system for visual identification of plant species. In *European Conference on Computer Vision*, pages 116–129. Springer, 2008.

[65] Carola Figueroa Flores, Abel Gonzalez-Garcia, Joost van de Weijer, and Bogdan Raducanu. Saliency for fine-grained object recognition in domains with scarce training data. *Pattern Recognition*, 2019.

[66] Qi Wang, JianJun Wang, Hongyu Deng, et al. Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification. *Pattern Recognition*, 2023.

[67] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1365–1372. IEEE, 2009.

[68] Lianbo Zhang, Shaoli Huang, and Wei Liu. Learning sequentially diversified representations for fine-grained categorization. *Pattern Recognition*, 2022.

[69] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[70] Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128, 2006.

[71] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019.

[72] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6599–6608, 2019.

[73] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[74] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

[75] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

[76] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

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

[78] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[79] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[80] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 541–549, 2017.

[81] Lukáš Picek, Milan Šulc, Jiří Matas, Jacob Heilmann-Clausen, Thomas S Jeppesen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020–not just another image recognition dataset. *arXiv preprint arXiv:2103.10107*, 2021.

[82] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[83] Kaiming He, Haoqi Fan, Yuxin Wu, et al. Momentum contrast for unsupervised visual representation learning. In *Computer vision and pattern recognition*, 2020.

[84] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, 2015.

[85] Yin Cui, Feng Zhou, Jiang Wang, et al. Kernel pooling for convolutional neural networks. In *Computer vision and pattern recognition*, 2017.

[86] Qilong Wang, Peihua Li, and Lei Zhang. G2denet: Global gaussian distribution embedding network and its application to visual recognition. In *Computer vision and pattern recognition*, 2017.

[87] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. 2017.

[88] Xing Wei, Yue Zhang, Yihong Gong, et al. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In *European Conference on Computer Vision*, 2018.

[89] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. 2018.

[90] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, et al. Pairwise confusion for fine-grained visual classification. In *European conference on computer vision*, 2018.

[91] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[92] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Computer vision and pattern recognition*, 2018.

[93] Zhihui Wang, Shijie Wang, Shuhui Yang, et al. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *Computer vision and pattern recognition*, 2020.

[94] Ruyi Ji, Longyin Wen, Libo Zhang, et al. Attention convolutional binary neural tree for fine-grained visual categorization. In *Computer vision and pattern recognition*, 2020.

[95] Adria Recasens, Petr Kellnhofer, Simon Stent, et al. Learning to zoom: a saliency-based sampling layer for neural networks. In *European conference on computer vision*, 2018.

[96] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13130–13137, 2020.

[97] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 2014.

[98] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[99] Lianbo Zhang, Shaoli Huang, and Wei Liu. Enhancing mixture-of-experts by leveraging attention for fine-grained recognition. *IEEE Transactions on Multimedia*, 2021.

[100] Lei Hu, Shaoli Huang, Shilei Wang, Wei Liu, and Jifeng Ning. *Do We Really Need Frame-by-Frame Annotation Datasets for Object Tracking?*, page 4949–4957. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475365.

[101] 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*, pages 7278–7286, 2018.

[102] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in neural information processing systems*, 2018.

[103] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9258–9267, 2019.

[104] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10657–10665, 2019.

[105] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Mahesh Prakash, and Alexander G Hauptmann. Few-shot activity recognition with cross-modal memory network. *Pattern Recognition*, 108:107348, 2020.

[106] Lunke Fei, Bob Zhang, Jie Wen, Shaohua Teng, Shuyi Li, and David Zhang. Jointly learning compact multi-view hash codes for few-shot fkp recognition. *Pattern Recognition*, 115:107894, 2021.

[107] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognition*, page 108111, 2021.

[108] Geonuk Kim, Hong-Gyu Jung, and Seong-Whan Lee. Spatial reasoning for few-shot object detection. *Pattern Recognition*, 120:108118, 2021.

[109] Jincheng Xu and Qingfeng Du. Learning transferable features in meta-learning for few-shot text classification. *Pattern Recognition Letters*, 135:271–278, 2020.

[110] Yikai Wang, Li Zhang, Yuan Yao, and Yanwei Fu. How to trust unlabeled data instance credibility inference for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[111] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern recognition*, 113:107826, 2021.

[112] Wei Zhu, Wenbin Li, Haofu Liao, and Jiebo Luo. Temperature network for few-shot learning with distribution-aware large-margin metric. *Pattern Recognition*, 112:107797, 2021.

[113] Yu Song and Changsheng Chen. Mppcanet: A feedforward learning strategy for few-shot image classification. *Pattern Recognition*, 113:107792, 2021.

[114] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[115] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[116] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.

[117] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[118] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[119] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[120] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[121] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[122] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, pages 3018–3027, 2017.

[123] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[124] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[125] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International Conference on Machine Learning*, pages 7115–7123. PMLR, 2019.

[126] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

[127] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9715–9724, 2019.

[128] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

[129] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[130] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[131] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611, 2006.

[132] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[133] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[134] Zhong Ji, Xingliang Chai, Yunlong Yu, and Zhongfei Zhang. Reweighting and information-guidance networks for few-shot learning. *Neurocomputing*, 423:13–23, 2021.

[135] Yiluan Guo and Ngai-Man Cheung. Attentive weights generation for few shot learning via information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13508, 2020.

[136] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[137] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[138] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. 2014.

[139] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision*, pages 329–344. Springer, 2014.

[140] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[141] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[142] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12576–12584, 2020.

[143] Kui Jia and Shaogang Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. In *IEEE International Conference on Computer Vision*, volume 2, pages 1683–1690. IEEE, 2005.

[144] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[145] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011.

[146] Simon Baker and Takeo Kanade. Hallucinating faces. In *IEEE international conference on automatic face and gesture recognition*, pages 83–88. IEEE, 2000.

[147] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[148] Wei Liu, Dahua Lin, and Xiaoou Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 478–484. IEEE, 2005.

[149] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE international conference on computer vision*, pages 4491–4500, 2017.

[150] Jae Young Choi, Yong Man Ro, and Konstantinos N Plataniotis. Color face recognition for degraded face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(5):1217–1230, 2009.

[151] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2019–2030, 2011.

[152] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.

[153] Bahadir K Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE transactions on image processing*, 12(5):597–606, 2003.

[154] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *IEEE conference on computer vision and pattern recognition*, pages 4792–4800, 2016.

[155] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[156] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *International Conference on Computer Vision*, pages 340–349, 2019.

[157] Xiaotong Zhao, Wei Li, Yifan Zhang, and Zhiyong Feng. Residual super-resolution single shot network for low-resolution object detection. *IEEE Access*, 6:47780–47793, 2018.

[158] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *Conference on Computer Vision and Pattern Recognition*, pages 14443–14453, 2021.

[159] Carlo Migel Bautista, Clifford Austin Dy, Miguel Iñigo Mañalac, Raphael Angelo Orbe, and Macario Cordel. Convolutional neural network for vehicle detection in low resolution traffic videos. In *IEEE Region 10 Symposium (TENSYMP)*, pages 277–281. IEEE, 2016.

[160] Xiao Wang, Jun Chen, Chao Liang, Chen Chen, Zheng Wang, and Ruimin Hu. Low-resolution pedestrian detection via a novel resolution-score discriminative surface. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1123–1128. IEEE, 2017.

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

[162] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[163] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2019.

[164] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6445–6449. IEEE, 2019.

[165] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[166] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[167] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[168] Guorui Zhou, Ying Fan, Runpeng Cui, Weijie Bian, Xiaoqiang Zhu, and Kun Gai. Rocket launching: A universal and efficient framework for training well-performing light net. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[169] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7028–7036, 2021.

[170] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[171] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *European Conference on Computer Vision (ECCV)*, pages 335–350, 2018.

[172] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[173] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[174] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1628–1636, 2021.

[175] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers

with localizable features. In *IEEE conference on computer vision and pattern recognition*, 2019.

[176] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.

[177] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[178] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019.

[179] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[180] Tian Zhang, Dongliang Chang, Zhanyu Ma, and Jun Guo. Progressive co-attention network for fine-grained visual classification. In *Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[181] Xiu-Shen Wei, Peng Wang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, 28(12):6116–6125, 2019.