SD-FSOD: Self-Distillation Paradigm via Distribution Calibration for Few-Shot Object Detection

Han Chen, Qi Wang, *Member, IEEE*, Kailin Xie, Liang Lei, Matthieu Gaetan Lin, Tian Lv, Yongjin Liu, *Senior Member, IEEE*, and Jiebo Luo, *Fellow, IEEE*

***Abstract*—Few-shot object detection (FSOD) aims to detect novel targets with only a few instances of the associated samples. Although combinations of distillation techniques and meta-learning paradigms have been acknowledged as the primary strategies for FSOD tasks, the existing distillation methods exhibit inherent biases and sensitivity to novel class variability. A critical hurdle for FSOD distillation is the difficulty in ensuring appropriate knowledge learned from the teacher model during the fine-tuning stage. Furthermore, coarse distillation procedures risk misalignment between the learned and actual distributions. This misalignment could potentially negate the benefits of positive cases and impede the detector's evolution. To address these deficiencies, we propose a novel self-distillation paradigm exclusively for the fine-tuning stage (SD-FSOD). Our methods integrate a Distribution Prototype Extractor (DPE) and Self-Distillation Memory (SDM), promoting feature distribution consistency during distillation. In detail, the DPE module reliably initializes the weights of the detector, ensuring a robust class distribution for the distillation process. Meanwhile, the SDM module utilizes decoupling techniques to divide the distillation tasks into two sub-task branches, allowing the student model to independently learn and share precise features through isolated distillation processes. The synergistic integration of feature calibration techniques and the continuous self-distillation paradigm distinctly enhances the fine-tuning process, which shows the superiority of the FSOD self-distillation methodologies. The extensive experiments on the PASCAL VOC and MS COCO datasets demonstrate that our proposed approach produces significant improvements and achieves state-of-the-art (SOTA) performance.**

***Index Terms*—Few-shot object detection, self-distillation, distribution prototype, decoupled sub-tasks.**

I. Introduction

Deep learning has made tremendous progress in the field of object detection [2], [44], which can be applied in autonomous vehicles, surveillance systems, and security, identifying and tracking objects of interest. However, one major criticism is the heavy reliance on large-scale annotated datasets, which are both resource- and time-intensive to acquire. Obtaining a sufficient amount of labeled data can often be challenging [31], and the lack of labeled data presents a substantial obstacle in real-world scenarios such as medical image analysis, deep-sea exploration, and rare object recognition. Few-shot object detection (FSOD), which aims to train an object detector that can generalize effectively with just a few numbers of annotated samples, has emerged as a solution to these problems.

Fine-tuning pre-trained models are the dominant approaches for few-shot object detection (FSOD). This process involves taking the detector pre-trained on a large dataset of base classes and then making minor adjustments to the model to adapt it to novel classes with limited data. Despite only incremental changes, fine-tuning has achieved impressive performance gains for FSOD by transferring knowledge from the base classes [6], [33], [39]. Building upon the fine-tuning approaches, several techniques have been developed, such as the fine-tuning of feature attention mechanisms, student-teacher distillation, and data augmentation strategies. Moreover, distillation-based approaches have gained prominence as highly effective and straightforward strategies to enhance FSOD detector performance. Additionally, distillation techniques provide an effective remedy for the challenges of knowledge retention and catastrophic forgetting that often plague fine-tuning methods [25], [36].

In general, the previous distillation-based strategies compare teacher-created class prototypes with student predictions for the same image in order to maximize the students' weights using a distillation loss [52]. For example, knowledge distillation is used to train teacher networks to produce prototypes of

Manuscript received 24 July 2023; revised 11 November 2023; accepted 11 December 2023. Date of publication 15 December 2023; date of current version 3 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62162008, in part by the Beijing Natural Science Foundation under Grant L222008, in part by the Guizhou Provincial Science and Technology Projects under Grant ZK[2022]-108, and in part by the Guizhou Overseas Educated Talents Innovation and Entrepreneurship Program under Grant 2022-04. This article was recommended by Associate Editor S. T. Kim. *(Corresponding authors: Qi Wang; Liang Lei.)*

Han Chen, Kailin Xie, and Liang Lei are with the Guangdong University of Technology, Guangzhou 510006, China (e-mail: 1711221608@qq.com; xiexyweizhi@163.com; leiliang@gdut.edu.cn).

Qi Wang is with the State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou 550025, China, and also with the College of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: qiwang@gzu.edu.cn).

Matthieu Gaetan Lin, Tian Lv, and Yongjin Liu are with the College of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: lin21@mails.tsinghua.edu.cn; t22@mails.tsinghua.edu.cn; liuyongjin@tsinghua.edu.cn).

Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: jiebo.luo@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3343397.

Digital Object Identifier 10.1109/TCSVT.2023.3343397

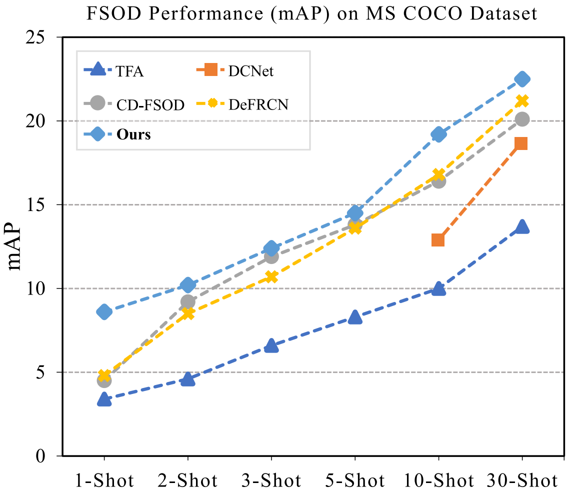
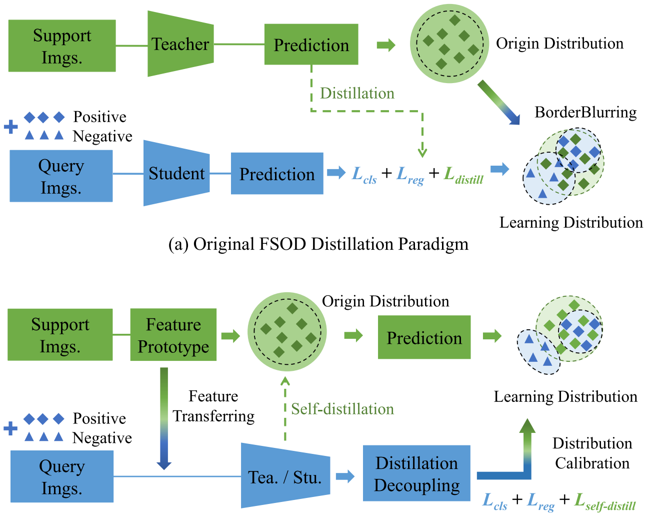


Fig. 1. FSOD performance (mAP) on MS COCO novel sets for K-Shot numbers. Our proposed SD-FSOD significantly outperforms previous SOTA methods.



(a) Original FSOD Distillation Paradigm

(b) Our FSOD Self-distillation Paradigm

Fig. 2. Comparative illustration of the FSOD distillation process. (a) Original FSOD distillation methods rely solely on support information to pre-determine the teacher modelpotentially resulting in ambiguous class distribution boundaries. (b) Our FSOD self-distillation paradigm dynamically refines itself with newly acquired knowledge based on the strong support of class prototypes. Additionally, our method incorporates distillation decoupling to mitigate distribution bias.

classes. Alternatively, with the help of the student network, the model uses fine-tuned convex optimization to learn the class features of the labels. In essence, the effectiveness of these FSOD distillation methods is dependent on the student model's capacity to extract critical features and construct a sufficiently robust feature distribution, even when only a limited number of samples are available.

Although the previous approaches have brought about further improvements, the distillation mode has an assortment of persistent issues: a) Acquiring the optimal distribution for the teacher model during the fine-tuning process still poses significant challenges [38]. b) Knowledge distillation may cause biases in confidence and precision, leading to ambiguous category heterogeneity [15]. c) The crude distillation approaches tend to ignore classification and orientation tasks that have different learning preferences [43], resulting in biased feature weights and suboptimal performance.

Specifically, previous research indicates that the distillation-based strategies for FSOD models are feasible and perform better than conventional fine-tuning schemes [32], [37]. As shown in Figure 1, DCNet [14] and CD-FSOD [45] apply the overall distillation scheme to bring considerable improvement to the fine-tuning paradigm model, verifying the important contribution of the distillation mode to few-shot scenarios. However, these two approaches primarily concentrate on searching a robust teacher model, which in practice remains a challenging and critical task in FSOD applications. A robust teacher model is not only difficult to identify, but the misalignment in its transfer process cannot be guaranteed. Knowledge self-distillation has emerged as an effective solution to the problem of feature bias [24]. To further mitigate the uncertainty caused by the quality of the teacher model, we propose a novel self-distillation paradigm that considers class prototype support and distillation hobbies, addressing the difficulty of finding a suitable role model to learn from. We analyze that establishing a strong prototype distribution is fundamental to the successful process of distillation. Additionally, we devise specialized classification and regression tasks based on decoupling technology to fully take advantage of self-distillation. This method adeptly resolves solve the distillation conflict between translation-invariant features of class-agnostic RPN and migration-covariant features of class-related RCNN. The experimental results demonstrate that our decoupling self-distillation approach with the support of class prototype distribution outperforms the prior state-of-the-art methods.

To express our design motivations more concretely, we demonstrate the differences between our distillation paradigm (SD-FSOD) and previous approaches, as shown in Figure 2. In contrast to the previous practices where the student models acquire knowledge from the teacher models, we apply the local self-distillation technique during the fine-tuning phase. First, we introduce a powerful class prototype construction technique designed to build a robust prototype distribution. Such a unique approach enables dynamic calibration of query features during the distillation process, preventing feature distribution bias resulting from over- or under-distillation. Second, by integrating soft targets into self-distillation subtasks, the detector heads utilize self-feedback to optimally adapt to inter-class category heterogeneity and facilitate the acquisition of intra-class feature consistency. Our construction methods can not only prevent the problem of fuzzy margins of feature distribution caused by wrong distillation but also cultivate the self-reinforcing and self-improving discriminative ability of features under the premise of correctly identifying the model. Essentially, we aim to make our model serve as its own teacher, fostering a process of self-evolution in the fine-tuning stage.

We conduct extensive experiments to validate our detection model's strong performance and robustness on both the

PASCAL VOC and MS COCO datasets. Our technical contributions can be summarized as follows:

- We investigate the limitations and potential of the distillation model in the FSOD scenarios and devise a more suitable self-distillation few-shot approach, SD-FSOD, which applies distillation only in the fine-tuning stage.
- We propose a scattering approach to extract a robust class prototype suitable for distillation, mitigating knowledge feature bias and allowing the model to better accommodate inter-class heterogeneity and absorb intra-class consistency in the distillation process.
- We decouple the distillation process into two subtasks, categorical distillation and locational distillation, to avoid interference and maintain a joint representation during fine-tuning.

The remainder of this paper is organized as follows: In Sec. II, we discuss the related work on few-shot object detection. In Sec. III, we present the overall framework and the proposed method in detail. Sec. IV reports the experimental and ablation results, and Sec. V concludes the paper.

II. Related Work

Our work is related to several fields, including few-shot learning, few-shot object detection, and knowledge self-distillation memory in the visual field. We review these related works in the following parts.

A. Few Shot Learning (FSL)

Due to its capacity to learn with little data, few-shot learning [53] has attracted a lot of interest in the domain transfer learning field. Three major paradigms can be used to categorize few-shot learning tactics: data augmentation strategies, meta-learning, and metric learning. a) Data enhancement strategies [19], [20], including image transformation, synthetic data generation, and pseudo-labeling, are employed to augment the sample size and enhance the model's generalization ability through image processing and the synthesis of additional data. b) Meta-learning-based methods optimize a particular model to acquire a learner that can adapt to new tasks, also known as "learning to learn" [29], [30], [34]. Meta-learning approaches primarily gauge the similarity between support set images and the test image, leveraging that category for model prediction. c) Metric learning strategies [5], [7] are derived offshoots of meta-learning. By establishing a distance metric to measure similarity between query and support samples, these methods can accurately categorize data even without large training datasets. Despite the effectiveness of the above methods for classification tasks, their application in complex few-shot object detection (FSOD) tasks presents challenges such as object occlusion, distribution confusion, and variable scales. Consequently, the development of effective FSOD methods that can overcome these challenges is a constantly evolving area of research [26].

B. Few Shot Object Detection (FSOD)

With only a limited number of annotated samples, few-shot object detection (FSOD) endeavors to identify distinct categories [3]. However, many FSOD methods struggle to identify novel, arbitrary unseen categories, as they often rely on artificially assigned pre-existing categories. Inspired by meta-learning, early FSOD methods utilize meta-learners to generalize feature weights to novel classes. For instance, Meta R-CNN [47] and FSRW [17] use support images to aggregate query features, leading to a variety of feature aggregation and spatial augmentation techniques. Among these methods, A-RPN [10] stands out because it makes use of attentional RPN to exclude background boxes and features from particular classes. The fine-tuning paradigm, which incorporates a two-stage training process to improve the quality of information transmission, is the leading strategy at the moment. TFA [39], for example, only has to tweak the final layer to produce passable results, and FSCE [33] makes use of supervised contrastive loss to overcome problems with insufficient inter-class matching. DeFRCN [27] introduces gradient decoupling technology that utilizes fine-tuning techniques to optimize classification and regression tasks in parallel. Additionally, certain semi-supervised learning approaches have demonstrated efficacy in FSOD tasks. For instance, LVIS [19] implements the enlargement of novel classes through pseudo-annotation. VAE [46] mitigates variation in sample distribution by generating features with increased clip-related diversity. However, simple methods for fine-tuning might run into issues including model knowledge loss, trouble adjusting to novel weights, and feature distribution shifts as new features are added. Consequently, we suggest a reliable fine-tuning framework dubbed the SD-FSOD (Dynamic Self-Distillation Framework) to improve the model's knowledge transfer ability.

C. Self-Distillation Mechanisms (SDM)

For few-shot object detection (FSOD), the proportion of each new sample in the feature weights is much larger than in regular object detection tasks. Several studies have explored the efficacy of distillation mechanisms in the context of FSOD tasks. EKD [35] introduces a progressive approach to knowledge distillation that enhances the efficacy of transferring knowledge from teacher models. Similarly, MFDC [43] applies a knowledge distillation memory bank to learn and extract feature commonalities between base classes and novel classes, which culminates in exceptional performance metrics. However, it is particularly difficult to find a suitable teacher model when the sample is insufficient, and crude learning methods can introduce bias in the distribution of features [45], [52]. Researchers have proposed the self-distillation methods, in which the student model acts as its own instructor to overcome this problem. The self-distillation strategies can hasten the transfer of knowledge while enhancing learning capacity [32], [45]. By fully exploring the feature correlations between base and novel classes, the self-distillation approaches allow for a more accurate representation of the data distribution, overcoming the limitations of the global distillation mode. Inspired by these works, we propose a novel self-distillation paradigm to accomplish soft knowledge distillation through a potent class prototype during the FSOD's fine-tuning step.

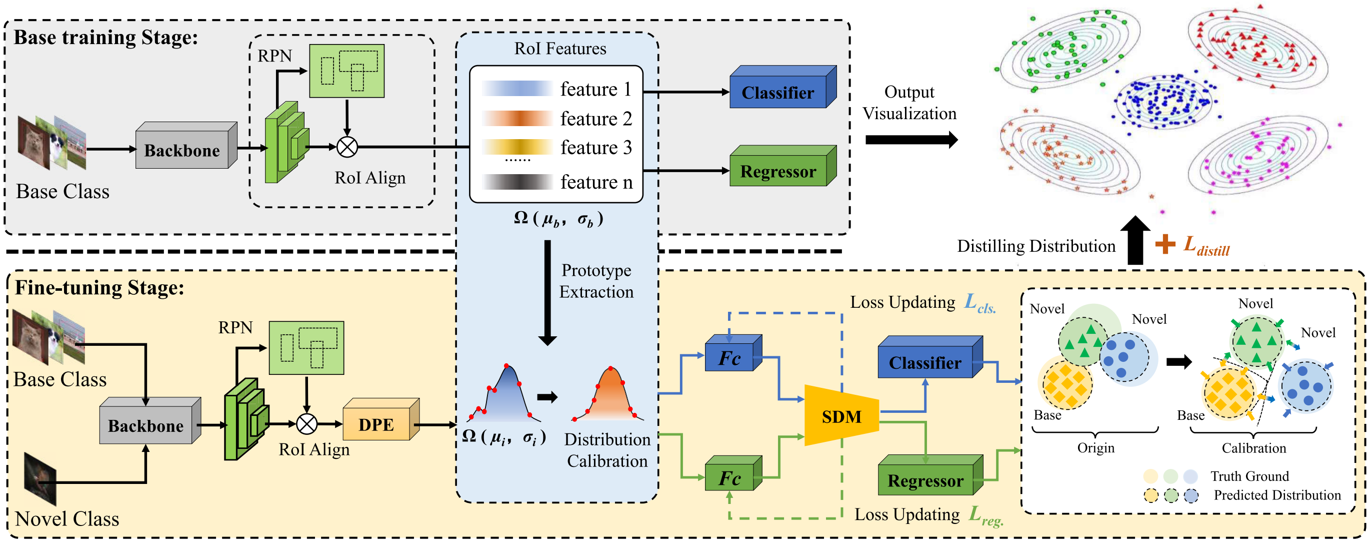


Fig. 3. The proposed framework of SD-FSOD. Compared to the standard Faster R-CNN, there are two additional components inserted into the framework: the Distribution Prototype Extractor (DPE) module and the Self-Distillation Memory (SDM) module. DPE applies base feature to initialize the detector's weights to ensure a robust distribution of novel classes. SDM separates the self-distillation process into two subtasks, which allows the student model to acquire and exchange precise features through independent distillation processes. (The blue line represents the distillation task of classification, and the green line represents the distillation task of location regression.)

D. Prototype Distribution Calibration (PDC)

Prototype distribution calibration (PDC) has become increasingly used in both semi-supervised and weakly supervised scenarios due to its outstanding ability to mine feature correlations [49]. Building on this strength, SCM [54] adopts a spatial calibration module to dynamically adjust semantic relevance and spatial context strength, which demonstrates the excellence of PDC in weakly supervised object localization. By creating an image-specific prototype, SIPE [4] captures complete regions and utilizes consistency loss to enhance segmentation performance. It can be concluded that a robust prototype distribution is crucial for the model's performance. However, FSOD models are extremely sensitive to the variations of novel classes during the fine-tuning stage, which can lead to significant deviations in distribution. Fittingly, distribution calibration technologies leverage feature correlations to guide the formulation of novel class feature distributions, consequently mitigating feature ambiguity. In line with this trend, TEDC [51] employs a distribution calibration module to reduce the deviation between the distributions of support features and query features in the same class. Inspired by the above excellent works, we cleverly integrate class prototype distribution calibration technology within the self-distillation paradigm for FSOD tasks. Our distribution prototype extractor effectively counters the challenges associated with the quality of the teacher model and fosters the model's ability to identify and filter the acquired knowledge in the subsequent distillation.

III. Methodology

The FSOD preliminary is covered in Sec. III-A. In Sec. III-B, and we illustrate the overall framework, Self-Distillation Paradigm via Distribution Calibration for Few-Shot Object Detection (SD-FSOD). Then, we discuss the Distribution Prototype Extractor (DPE) module and the Self-Distillation Memory (SDM) module in Secs. III-C and III-D.

A. Preliminary

We implement the two-stage training approaches following the standard FSOD setups [39], [47]. Let $D = \{(x_i, y_i)|x_i \in X, y_i \in Y\}$ denotes the training set, where x_i represents object images and $y_i = \{c_i, b_i\}$ represents the corresponding label consisting of class c_i and bounding box b_i. The classes are composed of base classes C_{base} with sufficient annotations and novel classes C_{novel} with only K (no more than 30) instances, and the two datasets are mutually exclusive ($C_{base} \cap C_{novel} = \emptyset$). In the first stage, we train a pre-trained detector using the abundant data from the C_{base}. In the second stage, we fine-tune the detector by jointly adding the novel classes to optimize a detector capable of detecting the novel classes. Finally, we evaluate the performance of the few-shot detector on $D_{test} \sim \{C_{base} \cap C_{novel}\}$.

B. Framework

We choose Faster R-CNN [28] as the foundational detector and further built our overall framework (SD-FSOD), as shown in Figure 3. The training process is divided into two stages, the base training stage and the fine-tuning stage. In the base training stage, we adopt the training methodology as well as the original RCNN model for object detection. During the fine-tuning stage, we present the Distribution Prototype Extractor (DPE) module during feature migration, which focuses on building a robust class prototype using correlation information between the base class features and the novel class features. We then propose the Self-Distillation Memory (SDM) module to devise two different self-distillation tasks for classification and regression. The purpose of SDM is to perform a distribution calibration by a self-supervised method, which is to

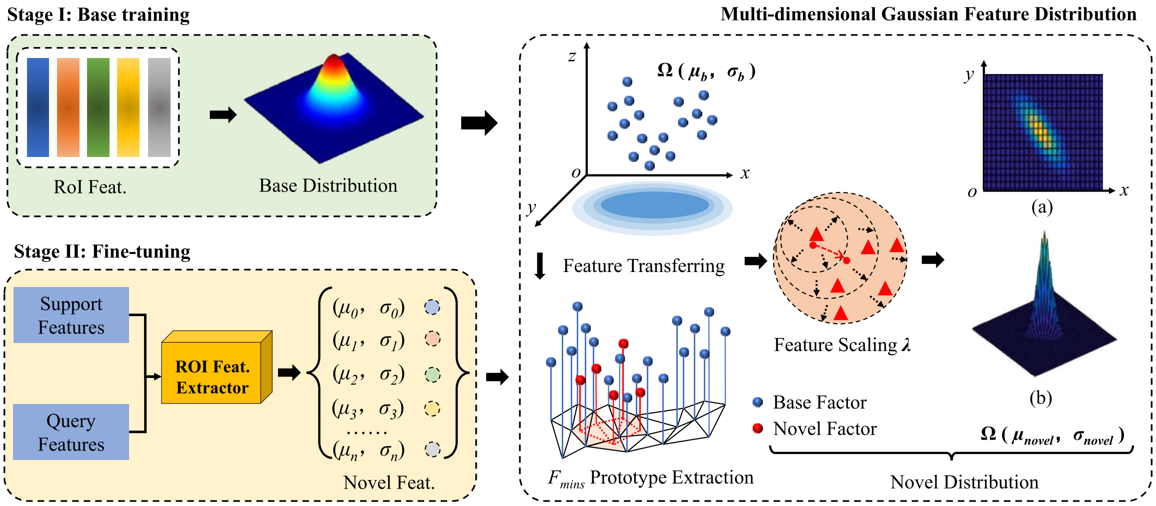


Fig. 4. The detailed structure of the proposed DPE module. In the base training stage, a relatively stable base distribution is obtained. During the fine-tuning stage, DPE extracts and calibrates scattered novel class features based on transferring feature knowledge and acquires a relatively robust novel distribution through gradient scaling.

predict the distribution closer to the true distribution. The framework provides a powerful and robust FSOD distillation mode, and we will detail its structure in the following sections.

C. Distribution Prototype Extractor (DPE)

The existing FSOD methods may struggle to ensure an optimal feature prototype distribution that capturing the distinctive characteristics of object classes with minimal training data remains a challenge. Class imbalance can affect the feature distribution, leading to biases and inaccuracies in the representation of different object classes [32], [45]. To mitigate the boundary ambiguity caused by class imbalance, we consider constructing feature distributions by class prototype calibration techniques. The distribution of feature prototypes plays a crucial role in capturing the discriminative characteristics of object classes. We propose the Distribution Prototype Extractor (DPE), whose primary purpose is to align the target distribution with its standard distribution and so provide a reliable feature representation for later usage. When encountering new objects, the model can leverage the learned feature prototypes to make accurate predictions based on their similarity to the prototypes, even without extensive training data for these novel classes.

As shown in Figure 4, the base class distribution is obtained through the base training stage. Then the DPE component is embedded in the RoI feature extractor to fine-tune training. Then, the feature extractor extracts the key feature points through the multidimensional Gaussian distribution and uses the mean and covariance of the features to fit a stable novel-class feature prototype, i.e. normalized prototype. Such a prototype can provide a good prerequisite for subsequent distillation knowledge acquisition, which can efficiently apply Gaussian offsets to sharpen the model's predicted values. In the following part, we describe these two components in detail.

1) RoI Feature Extraction: This component computes the class distribution from the received proposals, which is represented by the Gaussian distribution $\Omega_i = \{(\mu_i, \sigma_i)\}$, where μ_i and σ_i denote the mean and covariance of the class distribution, respectively. After the base training stage, a feature representation $\Omega_b = \{(\mu_b, \sigma_b)\}$ is obtained with specific formulas for the mean μ_b and covariance σ_b, as follows:

$$\mu_b = \frac{1}{N_{\mathrm{n}}} \sum_{i \in N_b} x_i^s, \tag{1}$$

$$\sigma_b = \frac{1}{N_n - 1} \sum_{i \in N_b} \left(x_i^s - \mu_b\right)\left(x_i^s - \mu_b\right)^T, \tag{2}$$

where N_n is the total number of classes, N_b is the particular base class, and x_i^s is the feature vector of the class. s is the training iteration steps.

2) Normalized Prototype: In the fine-tuning stage, we apply the normalization strategies to scale novel classes the elements in the feature tensor by adding an alignment compensation factor λ. The factor can reduce the differences in distribution to ensure that the base distribution does not mislead the construction of the novel distribution. The essence of the factor is to adjust the process of distribution construction by means of gradient scaling, which improves the stability of the model's performance. The construction process is as follows:

$$\mu = \lambda \times \mu_b + (1 - \lambda) \times x_i^s, \tag{3}$$

where μ is the mean of the distribution prototype. We analyze the value range of λ into three situations. When λ<0, it adversely affects the optimization trajectory for RoI, which is equivalent to the disappearance of the adversarial strategy during the feature migration process. When λ>1, the migration effect is overemphasized, and not considering the novel feature knowledge is not enough. Therefore, we initialize to a range interval: (0, 1]. We empirically set λ to 0.33 to normalize the scale of the distribution by default. The few feature factors that are most similar to the base class are used to determine the center of the novel class feature distribution after the normalized feature distribution has been obtained.

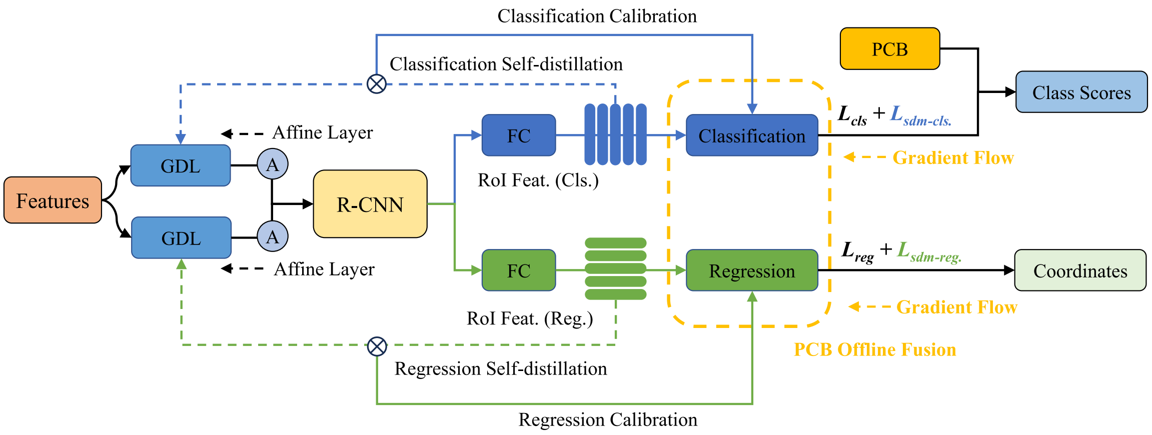


Fig. 5. The detailed structure of the proposed SDM module. GDL utilizes affine transformation strategies to modulate the decoupling degree between modules. In the process, the self-distillation task is decoupled into relatively independent subtasks on the FC layer. The output loss of the classification task and regression task will be calibrated and fused by the offline PCB. (The blue line is the distilling task of classification, and the green line is the distilling task of location regression.)

Through gradient computation, the building process of the novel distribution Ω_n is demonstrated below:

$$\mu_n = x_i + \frac{1}{N}\sum_{i\in n}^{|F_{mins}|} \mu, \tag{4}$$

$$\sigma_n = \frac{1}{k}\sum_{i\in n}^{|F_{mins}|} \mu_n + C, \tag{5}$$

where μ_n and σ_n denote the mean and covariance of the novel distribution, respectively. Furthermore, we focus on refining the feature commonality between the base classes and the novel classes during the fine-tuning stage. To mitigate the common interference between base classes, we compress the similarity covariance σ_b of the region proposals from one base class to another base classes, and σ_b is normalized as a small constant value C. $C \in \{\lambda\sigma_b/C_{gt}\}$. C_{gt} is the groundtruth distribution, which can preserve and refine the commonality between effective base classes without disturbing novel distribution. Additionally, this compression method concentrates on extracting common features between base and novel classes without adding extra computational load. x_i is the normalized feature vector, and k is the number of feature points that are closest to the center of the distribution. $|F_{mins}|$ are the feature factors that are most similar to the base class.

We primarily use a measure of the difference between adjacent distributions to gauge whether the distributions of data samples are similar. In object detection, differences in feature distributions may cause a model to perform poorly in different domains. Our distribution calculation allows the model to better learn the distribution of features, which in turn improves the generalization performance over differently distributed data. Our approach combines the benefits of preadaptation and category balancing to reduce the impact of distribution differences and encourage the model to learn consistent feature representations across samples. DPE effectively eliminates the confusing feature margin by intelligently leveraging center distances while ensuring the strong location of the feature center during the fitting of the novel class distribution, outputting a robust feature representation.

D. Self-Distillation Memory (SDM)

Distillation methods enhance the generalization capability of the student model to novel classes by leveraging the knowledge learned from a diverse set of base classes. In particular, the approaches help in improving the model's ability to recognize and classify objects from unseen classes during inference. However, dealing with intra-class variability, where objects within the same class may exhibit diverse appearances, remains a challenge in self-distillation. When the detectors are jointly optimized, RCNN demands translation-invariant features for box classifier whereas translation-covariant features for box regressor [27]. The previous models are still hard to capture and distill the essential characteristics of each class while accommodating the variations within the class [45]. Therefore, to solve this problem, we propose a Self-Distillation Memory (SDM) module that can adequately accommodate the changes in characteristics during the fine-tuning stage.

As shown in Figure 5, we adopt gradient decoupled layers (GDL) component and offline prototypical calibration block (PCB) component, which enhance feature representation and eliminate false positives for high scores, respectively [27]. At the same time, we also built our self-distillation architecture to prevent any implicit interference with the evolutionary process. Firstly, we apply the FC layer to extract the RoI feature matrix for mapping and distribute the feature vectors for knowledge self-distillation. Then, we use decoupling techniques, splitting the detection into two simultaneous subtasks and integrating innovative design elements. With the architecture, it is ensured that such interference does not affect the final loss joint output. Finally, we utilize an offline PCB module to fuse the distillation losses for classification and location calibration. The fusion technique is applied exclusively during the inference stage without any further training and can greatly improve the performance of few-shot detectors.

More specifically, such an approach can simultaneously compute the similarity in the pre-learned space instead of relying on the teacher model to make predictions for the categories. Therefore, the obtained region proposals can be used to calculate the similarity and serve as soft labels in the process of self-distillation to supervise the detector learning of the fine-tuning backbone task. It is worth noting that, unlike the previous practice of extensively distilling the entire framework, we only set up local self-distillation in the fine-tuning stage. In the following part, we illustrate the details of self-Distillation Memory (SDM) in four steps.

1) Feature Activation: The average feature after pooling random input is obtained from DPE as a memory sample, and its distribution Ω_d is expressed as:

$$\Omega_d = \left\{\Omega_p \left|\Omega_p \sim N(\mu_d, \sigma_d)\right.\right\}, \tag{6}$$

where Ω_p is the feature distribution after RoI Align pooling. More specifically, the newly selected query samples V_q are calibrated and estimated using the support samples V_s accumulated in the memory bank. The covariance σ_d is obtained by cosine distance, which can be specifically expressed as:

$$\sigma_d = \frac{1}{|k|}\sum_{i\in I}\frac{V_s^T V_q}{\left\|V_s^T\right\|\left\|V_q\right\|}, \tag{7}$$

where k is the number of random input samples. As the detector processes both query and support features, meta-feature categories and localization information for new objects will be reactivated.

2) Distillation Decoupling: These activation features are then used to perform distillation evolution. However, there are differences in the feature learning of the detector for the classification task and the localization task during the distillation learning process. Therefore, we decouple the self-distillation task into two adaptive branch sub-tasks. The classification and localization tasks are explicitly decomposed by using the fully connected FC layer. The two terminal tasks of the distillation process are decoupled into two independent learning spaces. We mainly aim to solve the problem that different tasks have different learning preferences for features of commonality in the distillation process. Therefore, its decomposition process is as follows:

$$\left\{R_i^{cls.}, R_i^{reg.}\right\}_{i=1}^{n} = F\left(\Gamma(D(Q))\right), \tag{8}$$

where $R_i^{cls.}$ and $R_i^{reg.}$ represent two decoupled sub-tasks, F is the FC layer, Γ is the RoI Align pooling, and $D(Q)$ represents the distillation task during the query process.

In the fine-tuning stage, self-calibration is performed alongside feature learning, and a dynamic displacement compensation process is performed to maximize the potential of the distilled model. The process is expressed as:

$$L_{sdm} = \frac{1}{\left|N_{\Omega_d}\right|}\sum_{v\in\Omega_d} L_{fine-tuning}(f(R_i)), \tag{9}$$

where distribution representation is denoted by the losses L_{sdm}, and $L_{fine-tuning}$ is the losses durling fine-tuning stage. N_{Ω_d} is the number of the distribution, v is the feature vectors of the distribution and $f(R_i)$ is the distillation subtask.

3) Self-Distillation Calibration: For the distillation end of the classification task, a cascade detector is used to process the target features step by step. At the same time, the distillation scheme combined with the bionic layer GDL and PCB connectivity enables the model to obtain learning features at a deeper level, avoiding the gradient disappearance and gradient explosion problems and helping to mine difficult sample features. Finally, the cross-entropy loss $L_{cross-entropy}$ calculation is used as follows:

$$L_{sdm-cls} = \frac{1}{\left|N_{\Omega_d}\right|}\sum_{v\in\Omega_d} L_{cross-entry} \times \left(-\frac{1}{N_i}\sum_{i\in I}\log\left(\frac{e^{s_i}\cdot e^{q_i}/\tau}{\sum\limits_{i\in I}(e^{s_i}\cdot e^{q_i})}\right)\right), \tag{10}$$

where $L_{sdm-cls}$ is the loss of the classification distillation. N_i is the input of different cascades, and e^{s_i} and e^{q_i} are feature vectors of the support set and of the query set, respectively.

For the distillation end of the regression task, a nonlinear mapping is used to initialize and normalize the positioning, and the class information is re-aggregated to form a bounding box representation of the class. In other words, the backbone positioning task is used as the main output, and the positioning information learned by distillation is used as an offset to minimize the gap between the predicted positioning value and the real position value, which is optimized by Smooth-$L1$ loss, as follows:

$$L_{sdm-reg} = \frac{1}{\left|N_{\Omega_d}\right|}\sum_{v\in\Omega_d} D(Q)\Big(\sum_{i\in\{x,y,w,h\}} Smooth_{L1}(g_i - p_i)\Big), \tag{11}$$

where $L_{sdm-reg}$ is the loss of distillation of the classification. g_i is the truth ground, p_i is the predition value. $\{x, y, w, h\}$ is the location of the precision.

4) Loss Function: Totally, the classification and bounding box regression losses of the RPN and the RCNN are included in the loss function, along with the losses incurred during self-distillation. The loss function L_{total} is jointly optimized as follows:

$$L_{total} = L_{RPN} + L_{cls} + L_{reg} + \lambda(L_{sdm-cls} + L_{sdm-reg}), \tag{12}$$

where L_{RPN} is the loss of RPN, λ represents the loss factor for the corresponding task, which is used to scale the distillation loss to fit the detection terminal task.

IV. Experiments

We introduce the experimental benchmarks in Sec. I. Then, in Sec. II, we provide a description of our implementation and analysis of the PASCAL VOC datasets and COCO datasets. Finally, in Sec. III, we present the qualitative findings of our ablation studies on PASCAL VOC datasets.

A. Datasets and Setups

1) Datasets: We follow the standard data settings and evaluation protocols from previous works by TFA and conduct extensive tests on our framework using the Pascal VOC 2007 [9], VOC 2012 [8], and MS COCO datasets [22] to assess its performance on the FSOD task. For PASCAL VOC, we set up three different splits [39], dividing its 20 categories into 15 base classes with rich annotations and 5 novel classes with

TABLE I

EXPERIMENTAL FSOD RESULTS ON THE PASCAL VOC DATASET. WE EVALUATE SD-FSOD PERFORMANCE (nAP50) ON THREE DIFFERENT SPLITS. RED/BLUE INDICATE SOTA/THE SECOND BEST. THE RESULTS ARE AVERAGED OVER MULTIPLE RUNS

Method / Shots	Publications	Novel Split 1					Novel Split 2					Novel Split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW [17]	*ICCV 2019*	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet [40]	*ICCV 2019*	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [47]	*ICCV 2019*	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
RepMet [18]	*CVPR 2019*	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
NP-RepMet [48]	*NeurIPS 2020*	37.8	40.3	41.7	47.3	49.4	41.6	43.0	43.4	47.4	49.1	33.3	38.0	39.8	41.5	44.8
TFA w/cos [39]	*ICML 2020*	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [42]	*ECCV 2020*	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
FSCE [33]	*CVPR 2021*	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
SRR-FSD [55]	*CVPR 2021*	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
CME [21]	*CVPR 2021*	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
Dictionary	*NeurIPS 2021*	46.1	43.5	48.9	60.0	61.7	25.6	29.9	44.8	47.5	48.2	39.5	45.4	48.9	53.9	56.9
FADI [1]	*NeurIPS 2021*	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6
UP-FSOD [41]	*ICCV 2021*	43.8	47.8	50.3	55.4	61.7	31.2	30.5	41.2	42.2	48.3	35.5	39.7	43.9	50.6	53.3
QA-FewDet [11]	*ICCV 2021*	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5
Meta-DETR [50]	*arXiv 2021*	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6
DeFRCN [27]	*ICCV 2021*	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4
LVIS [19]	*ICVF 2022*	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6
FCT [12]	*CVPR 2022*	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7
MFDC [43]	*ECCV 2022*	**63.4**	**66.3**	**67.7**	**69.4**	**68.1**	**42.1**	46.5	**53.4**	**55.3**	**53.8**	56.1	58.3	59.0	62.2	63.7
MFE [16]	*WACV 2023*	55.0	55.5	59.2	-	59.7	34.7	38.2	44.1	-	46.4	49.5	44.2	47.3	-	55.4
VFA [13]	*arXiv 2023*	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9
Vanila-VAE [46]	*CVPR 2023*	60.0	63.6	66.3	68.3	67.1	39.3	46.2	52.7	53.5	53.4	56.0	58.8	57.1	62.6	63.6
Norm-VAE [46]	*CVPR 2023*	62.1	64.9	**67.8**	**69.2**	67.5	39.9	**46.8**	**54.4**	54.2	53.6	**58.2**	**60.3**	**61.0**	**64.0**	**65.5**
Ours	/	**64.6**	**67.1**	67.4	69.0	**70.7**	**42.4**	**48.3**	52.7	**55.4**	**56.0**	**57.0**	**59.7**	**60.4**	**63.5**	**64.6**

only K instances ($K = 1, 2, 3, 5, 10$). We report the average precision at IoU = 0.5 of novel classes (nAP) and part of base classes (bAP) on the VOC 2007 test set. Regarding the MS COCO dataset [39], which consists of 80 categories, the same 20 categories as in VOC are used as novel classes, and the remaining 60 categories are also used as novel classes. For MS COCO, we have $K = \{10, 30\}$ settings and report the average precision at IoU = 0.5:0.95 of novel classes (AP) and (AP75).

2) Setups: To establish our framework, we employ DeFRCN as the baseline and incorporated Faster R-CNN as the base detector and ResNet-FPN as the backbone. We design the model using the PyTorch framework, with model hyperparameters from the detectron2 library set as the default parameters. In the experiment, the SGD optimizer with momentum is set to 0.9, and the batch size is set to 8. In addition, We set the initial learning rate to 0.01 for base training and 0.005 for fine-tuning, with a weight decay of 1e-4. All experiments are conducted using two RTX 3090 GPUs.

B. Performance

1) Results on PASCAL VOC: To illustrate the effectiveness of our method, we evaluate our proposed method and compared it with the state-of-the-art (SOTA) methods, as shown in Table I. Our method outperforms the other methods significantly, achieving better results than the other methods. First, in the 10-shot setting, AP50 is improved by 1.9% on average, and all three split sets achieve the best detection results, reflecting the stable gains of our method. Then, our method is also excellent for the detection of very low shots. For example, in the 1-shot and 2-shot settings, our detection results exceed 0.8% and 1.4% of the previous best results, respectively. Finally, the three split sets improved by 4.4%, 3.7% and 5.0%, respectively. Our method achieves performance nearly equivalent to that of Norm-VAE. While Norm-VAE primarily focuses on modulating the output generation model according to various attributes, our SD-FSOD is more concerned with how to fully explore the potential relationships of existing features. To a certain extent, MFDC outperforms some semi-supervised methods with variational aggregation or generative paradigms. This superior performance can be attributed to MFDC's capability to continuously refresh the feature distribution via a distilled memory bank. In contrast, the advantage of our SD-FSOD lies in its semi-supervised adaptability. We utilize a simpler self-query technique to formulate a self-distillation distribution rather than deliberately selecting fixed novel data as the original novel distribution. The all-around substantial performance improvement also shows that the variance of the model in few-shot detection is more stable and robust.

In addition, in order to comprehensively demonstrate the performance of SD-FSOD, we list the average detection performance of several base classes and all novel classes in Split 1 in Table II. We report the detection value nAP50 of the novel classes and list the detection mean value and the detection value mAP50 of the above classes. On the one hand, the results show that both the detection of the novel class and the base class has achieved excellent results, which implies that our framework has no specific type preferences and is more conducive to the promotion of FSOD scenarios. On the other hand, from the 3-shot and 10-shot results, it can be seen that our model leaves enough room for optimization for the learning of new classes, thus acquiring feature representations in a more comprehensive way.

TABLE II

Comparative Performance Analysis of Base Classes and Novel Classes. AP50 (%) of the 3/10 Shots on the PASCAL VOC Dataset. Our Method Demonstrates Consistent Improvements Over Nearly All Established Baselines

Shots	Methods	Novel Classes						Base Classes								mAP50
		Bird	Bus	Cow	Motorbike	Sofa	Mean	Aeroplane	Bicycle	Boat	Bottle	Car	Cat	Chair	Mean	
3	LSTD [3]	23.1	22.6	15.9	0.4	0.2	12.4	74.8	68.7	57.1	44.1	78.0	83.4	46.0	64.6	52.8
	FSRW [17]	22.3	19.1	40.7	20.4	27.0	26.7	73.6	73.1	56.7	41.6	76.1	78.7	42.6	63.2	55.2
	Meta R-CNN [47]	31.1	44.6	50.8	38.8	10.7	35.0	67.6	70.5	59.8	50.0	75.7	81.4	44.9	64.3	57.3
	AFD-Net [23]	51.8	60.3	43.8	60.5	31.3	49.5	69.9	75.2	56.9	57.9	79.5	84.2	47.9	67.4	63.6
	DeFRCN [27]	56.2	**79.0**	**69.7**	66.7	50.0	64.3	83.2	84.3	70.3	72.6	86.5	85.3	67.0	78.5	73.7
	Ours	**69.8**	77.6	58.3	**74.5**	**56.8**	**67.4**	**83.9**	**86.0**	**70.9**	**73.0**	**87.5**	**85.4**	**67.5**	**79.2**	**74.4**
10	LSTD [3]	22.8	52.5	31.3	45.6	40.3	38.5	70.9	71.3	59.8	41.1	77.1	81.9	45.1	63.9	59.4
	FSRW [17]	30.0	62.7	43.2	60.6	39.6	47.2	65.3	73.5	54.7	39.5	75.7	81.1	35.3	60.7	59.5
	Meta R-CNN [47]	52.5	55.9	52.7	54.6	41.6	51.5	68.1	73.9	59.8	54.2	80.1	82.9	48.8	66.8	63.8
	AFD-Net [23]	62.7	67.9	60.2	68.1	42.7	60.3	73.0	77.9	60.4	59.7	81.8	85.4	51.0	69.9	68.3
	DeFRCN [27]	52.0	79.5	73.1	72.1	**58.3**	67.0	83.0	84.6	71.5	74.0	87.6	86.5	64.4	78.8	74.4
	Ours	**64.0**	**80.3**	**78.3**	**74.3**	56.6	**70.7**	**84.0**	**84.8**	**72.8**	**74.1**	**88.5**	**87.5**	**66.2**	**79.7**	**77.0**

TABLE III

Few-Shot Object Detection Performance on Novel Classes on the MS COCO 10/30-Shot Tasks. We Report nAP/nAP75 (%) Performance on the 20 Novel Classes of MS COCO in the FSOD Setting. The Best Is in Bold

Method / Shots	Publications	10		30	
		nAP	nAP75	nAP	nAP75
FSRW [17]	*ICCV 2019*	5.6	4.6	9.1	7.6
MetaDet [40]	*ICCV 2019*	7.1	6.1	11.3	8.1
Meta R-CNN [47]	*ICCV 2019*	8.7	6.6	12.4	10.8
TFA w/cos [39]	*ICML 2020*	10.0	9.3	13.7	13.4
MPSR [42]	*ECCV 2020*	9.8	9.7	14.1	14.2
QA-FewDet [11]	*ICCV 2021*	11.6	9.8	16.5	15.5
SRR-FSD [55]	*CVPR 2021*	11.3	9.8	14.7	13.5
FSCE [33]	*CVPR 2021*	11.9	10.5	16.4	16.2
UP-FSOD [41]	*ICCV 2021*	11.0	10.7	15.6	15.7
CME [21]	*CVPR 2021*	15.1	16.4	16.9	17.8
DeFRCN [27]	*CVPR 2021*	18.6	17.6	22.4	22.2
Vanila-VAE [46]	*CVPR 2023*	18.7	17.6	**22.5**	22.2
Norm-VAE [46]	*CVPR 2023*	18.7	17.8	**22.5**	22.4
Ours	\	**19.2**	**20.0**	**22.5**	**23.3**

TABLE IV

Ablation Experiments on DPE on PASCAL VOC Split 1. The DPE Module Is Embedded During the Base Training Stage and Fine-Tuning Stage, Respectively. We Further Demonstrate the Efficacy of DPE in Various K-Shot Scenarios. The Results Are Averages of Multiple Runs

	Ablation	Split 1				
		1	2	3	5	10
Base -Training	Baseline	76.4	77.0	77.3	78.2	78.4
	Baseline + DPE	**78.2**	**78.1**	**78.0**	**78.5**	**79.0**
	Improvement ↑	**+ 1.8**	**+ 1.1**	**+ 0.7**	**+ 0.3**	**+ 0.6**
Fine-tuning	Baseline	57.0	58.6	64.3	67.8	67.0
	Baseline + DPE	**63.8**	**64.4**	**67.1**	**68.7**	**68.3**
	Improvement ↑	**+ 6.8**	**+ 5.8**	**+ 2.8**	**+ 0.9**	**+ 1.3**

2) Results on MS COCO: Compared to the VOC dataset, MS COCO's category information is more complex, and the detection circumstances are more difficult. As a result, we report the detection results of nAP and nAP75 and set the detection model's performance in 10-shot and 30-shot scenarios, respectively. As can be seen in Table III, our suggested strategy outperforms current FSOD techniques, produces results that are competitive, and yields a significant improvement, which proves that our model performs well in more complex and difficult detection scenarios. Among them, nAP75 is widely regarded as a more difficult evaluation index that is used to assess the model's performance and generalizability while dealing with a small number of samples. Our method outperforms the best method by about 3.4% and 1.0% in the 10-shot and 30-shot settings, respectively.

C. Ablation Study

To verify the effectiveness of the proposed module, we conduct the ablation experiments and analyzed the experimental results. In this part, we apply the comprehensive ablation studies to split 1 of PASCAL VOC. The experiments are conducted over 10 random runs, and the results verify the reasonability of SD-FSOD.

1) Ablation for DPE: To further explore the performance of DPE, we integrate the Distributed Prototype Extractor (DPE) module in the base training stage and fine-tuning stage, respectively. The effects of DPE on k-shot detection are evaluated and detailed in Table IV. First, compared with the baseline, the detection rate is improved no matter whether the DPE component is added in the base training stage or the fine-tuning stage. Notably, the improvement is more significant in the fine-tuning stage than in the base training stage because the created base distribution is stable enough after sufficient training. Second, DPE concentrates more on how to filter, extract, and calibrate knowledge from existing distributions during the fine-tuning stage. Consequently, DPE's utility is primarily manifested in generating novel distributions from the established base distributions. Among them, the improvement in the extremely low shot scenarios is particularly obvious, especially in 1 and 2-shot, which is increased by 6.8% and 5.8%, respectively. It demonstrates the ability of our module to effectively by using critical features in a limited sample for the prototype construction. Then, with the increase in the number of samples, the feature distribution is more consolidated, and the improvement of the model by DPE is still obvious. Finally, the detection rate increased by 1.3% in the 10-shot setting, demonstrating the DPE's ability to remain

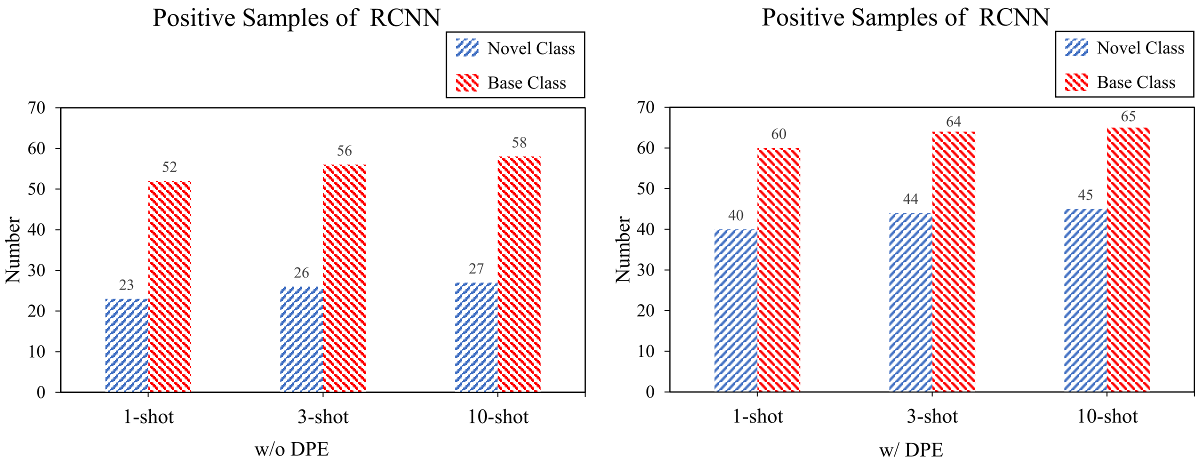


Fig. 6. Comparisons of the number of positive samples in RCNN. We investigate the performance of RCNN in extracting positive proposal boxes without(w/o) and with(w/) embedding the DPE module. DPE's feature mining capabilities are demonstrated for both base and novel classes.

TABLE V

ABLATION EXPERIMENTS OF SDM ON PASCAL VOC SPLIT 1. WE COMPARE THE PERFORMANCE SEVERAL METHODOLOGIES: THE TRADITIONAL FINE-TUNING TECHNIQUE, THE STANDARD DISTILLATION PARADIGM, SDM ENSEMBLE SELF-DISTILLATION, AND SDM DECOUPLING. THE RESULTS ARE AVERAGES OF MULTIPLE RUNS

No.	Meta-loss	Distillation	SDM	SDM Decouple		nAP50		
			Lsdm	*Lsdm-cls.*	*Lsdm-reg.*	1	5	10
1	✓					57.0	67.8	67.0
2	✓	✓				59.8	68.8	67.6
3	✓		✓			63.0	69.0	68.1
4	✓		✓	✓		**64.3**	68.5	69.5
5	✓		✓		✓	63.2	68.7	68.3
6	✓		✓	✓	✓	64.0	**69.0**	**70.3**

robust during scenarios with increased disturbances. In total, the experimental results show that DPE provides a powerful feature distribution prototype, which lays the foundation for the subsequent model operation.

In order to test the adaptability of the DPE and RCNN frameworks, we further conduct comparative experiments. We adopt a set of query samples to input randomly for detecting whether RCNN works with DPE, as shown in Figure 6. On the one hand, the number of positive samples detected after adding DPE has increased significantly, demonstrating that our module addresses two frequent issues with RCNN in few-shot detection and enhances the performance of RPN and RoI in the RCNN framework. On the other hand, this way reduces the false positives brought on by RPN's classification of decision-making as being too arbitrary when samples are sparse and also aids RoI in providing efficient feature representation under the assumption of prioritizing samples.

2) Ablation for SDM: To verify the performance of SDM, we design an ablation experiment to compare the baseline, the overall distillation scheme, the self-distillation mechanism, and the role of decoupling subtasks. In this experiment, we mainly report the nAP50 of various schemes as the detection standard, and the results are shown in Table V. First, we demonstrate the results of the baseline in No .1. Undoubtedly, compared with No. 1 and No. 2, the ensemble distillation plan enhances the paradigm fine-tuning approach. Second, as shown in No. 3, the improvement effect is very obvious, including a 4.8% improvement in 1-shot settings, in which our fine-tuned self-distillation framework is better suited to provide the correct guidance for the model. Then, we explore the performance of decoupling the SDM module into classification and regression subtasks, respectively. More specifically, we meticulously conduct a detailed assessment of the decoupling performance of each subtask. We mainly conduct comparative experiments by freezing one of the distilling subtask scores. Comparative experiments reveal that conducting either the classification or the regression sub-distillation task independently contributes to a modest enhancement in the model's overall performance. However, the model performance is not stable enough, especially for regression tasks that lack categorical information

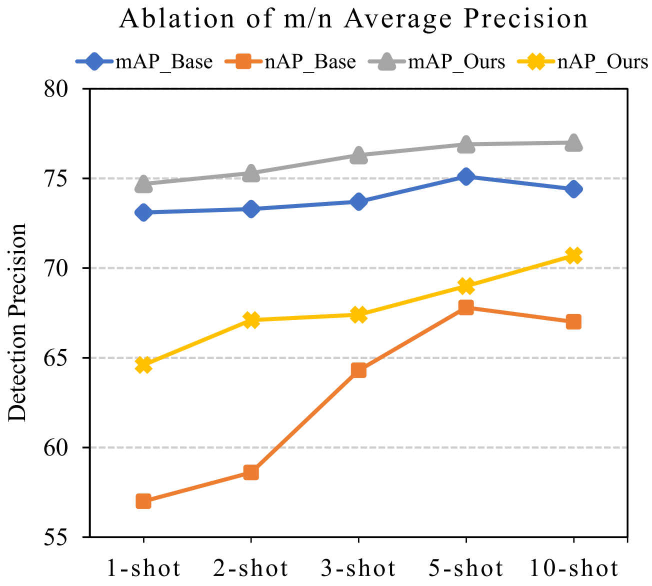


Fig. 7. Ablation on SD-FSOD of m/n Average Precision. We assess the synergistic effect of the DPE and SDM modules with a primary emphasis on mAP and nAP metrics. The detection accuracy comparison of K-Shot intuitively shows that our method far exceeds the baseline.

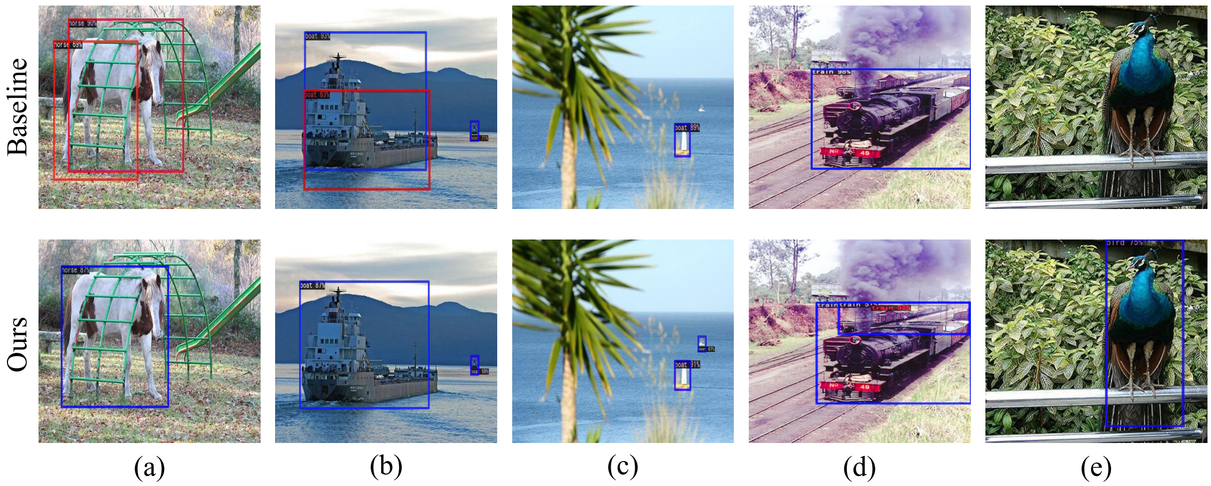


Fig. 8. Visualization of 1-shot object detection on the VOC dataset. We demonstrate the detection performance in different scenarios such as object occlusion, varying scales, small targets, and foreground-background confusion, respectively. We display bounding boxes with scores greater than 0.4. The success cases (blue boxes) and the failure cases (red boxes) are shown, respectively.

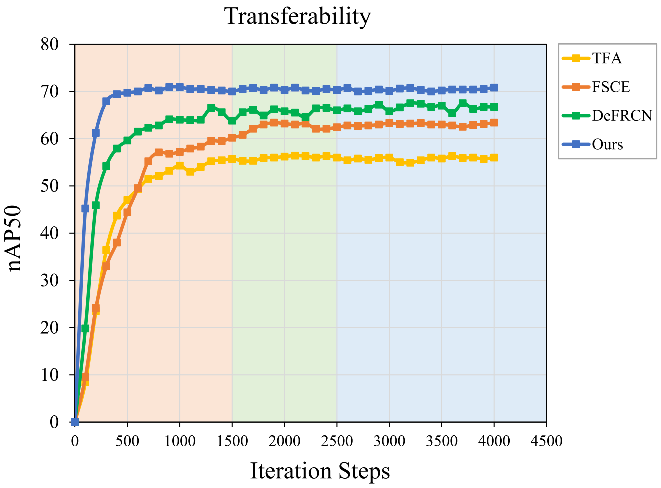


Fig. 9. Training speed comparison. We plot the change in nAP50 across training iterations. We compare the model evolution capabilities of the classic fine-tuning paradigm model at the same number of iteration steps.

feedback. Finally, we combine the loss scores from the two sub-tasks to jointly evaluate the distillation performance. Following our modified self-distillation decoupling technique, a thorough improvement and the best detection findings were obtained, as shown in No. 6. Overall, the decoupling technology in SDM distillation significantly enhances the baseline performance.

3) Ablation for SD-FSOD: To test the joint effect of class prototype extraction and self-distillation modality, we perform further ablation experiments. We adopt mAP50 and nAP50 as detection criteria, as shown in Figure 7. It is worth mentioning that mAP50 refers to the mean average precision when the recall rate is 50%, which mainly reflects the accuracy of the algorithm when detecting half of the positive samples. In general, a higher mAP50 means better performance of the algorithm in the FSOD task. First, whether it is mAP50 or nAP50, our method is much superior to the baseline, especially for the detection of various k-shot settings. Then, our nAP50 is more stable and less impacted by the number of samples, demonstrating the resilience and robustness of the model. As a result, the experiment demonstrates that a powerful class prototype is helpful in the construction of the distillation model and may also be employed as a teacher by itself to direct students' learning, which leads to amazing results.

D. Transferability

As shown in Figure 9, we evaluate the growth rate of nAP50 with the number of training iterations in the same operating mechanism to verify the transferability of our model. To ensure a level playing field for comparing and evaluating the performance, we eliminate potential biases introduced by variations in hardware performance by maintaining consistent computing equipment conditions and calculating the average value repeatedly. Furthermore, we employ the method of computing the average value repeatedly, allowing us to record the model's mean nAP50 detection rate at each hundred iteration step. First, the experiment shows that previous models are generally stable after the number of iterations reaches 2500, instead of our model almost entirely converging after 1500 iterations. Additionally, our model converges substantially quicker than previous SOTA methods for the same training iteration steps, indicating that our method has a faster rate of transferability and adaptation. Therefore, our method helps mitigate the impact of random fluctuations and provides a more reliable estimate of the model's detection.

E. Visual Inspections and Qualitative Results

1) Qualitative Visualization: As shown in Figure 8, we utilize visualization techniques to present the detection results of both the baseline model and our SD-FSOD network. The results clearly illustrate the notable improvements achieved by our method, addressing several challenges that the baselines have yet to overcome. First, through visual inspection of the detection results of Figure 8. (a) and (d), we can clearly observe the advantages of our SD-FSOD network over the

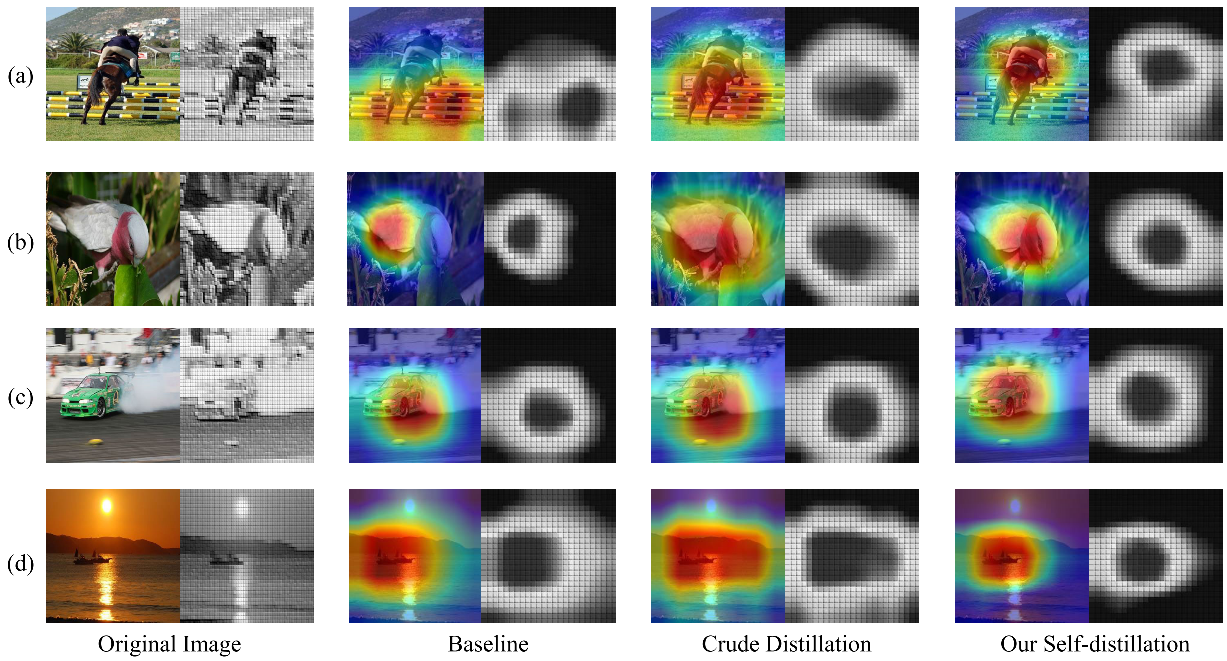


Fig. 10. Visualization of the heat map on the VOC dataset. We visualize the performance of conventional distillation methods and our self-distillation method in 1-shot scenarios. The results in each row demonstrate that our SD-FSOD has made significant improvements in solving the misalignment problem in distillation.

baseline model, such as the detection of the object occlusions and complex backgrounds. Second, it can be seen from Figure 8. (b) and (c) that our method shows greater capability in handling the detection tasks with multi-scale blending as well as fine targets. Finally, as shown in Figure 8. (a) and (b), our model improves the ability to reduce location bias, which shows the potential of knowledge distillation and fine-tuning processes in our SD-FSOD framework. In summary, our model is able to cope well with a variety of complex scenarios and has achieved excellent results.

2) Heat Map Visualization: To further validate the effectiveness of SDM, we randomly select some samples of the PASCAL VOC for testing in the 1-shot setting. And then we visualize the heat map of the feature information obtained by the detector, as shown in Figure 10. Specifically, we present the detection results of baseline, original ensemble crude distillation, and our self-distillation paradigm in four different complex scenarios, respectively. First, it's evident that the baseline model struggles in scenarios with target overlap and foreground confusion, showcasing limited capacity in capturing effective feature information. While the crude distillation method has enhanced feature extraction, the inferred results indicate that it primarily expands the global feature search capabilities. This approach may occasionally introduce more pronounced biases, as illustrated in Figure 10. (c) and (d). In contrast, our method demonstrates a superior ability to mine more comprehensive feature information, focusing on more regions and fine-grained features. The diverse results obtained from these complex scenarios conclusively illustrate our model's robust local and global feature search proficiency. Our method effectively addresses the misalignment issue in distillation and mitigates the problem of distillation flooding caused by limited sample. Thus, the SD-FSOD network has an excellent ability to address the crucial challenges in few-shot object detection.

V. Conclusion

In this paper, we address the critical issue of the distributional misalignment of class prototypes and the limitations of distillation frameworks in few-shot object detection (FSOD). Through comprehensive exploration of category relationships in FSOD scenarios, we introduce a powerful class prototype self-distillation network framework (SD-FSOD). Our proposed approach boasts several key contributions. First, we design the DPE module that generates a robust representation to effectively construct feature distributions, mitigating the issue of class prototype misplacement. Next, we utilize the SDM module with decoupling strategies to enhance learning preferences in both classification and regression tasks. Our proposed approach demonstrates the advancement of knowledge self-distillation in FSOD tasks. It opens up new directions for further study by fusing the class prototyping approach with the self-distillation paradigm, motivating the model to self-calibrate during the fine-tuning process. Experimental results on the PASCAL VOC and MS COCO datasets demonstrate the significant enhancements achieved by our proposed method. While our method has shown promising results, it is essential that it still requires further improvement, such as model adaptation, dependence on prototypes, and distillation efficiency issues. In conclusion, our approach significantly contributes to the innovations of FSOD research. We encourage researchers to build upon our ideas to drive progress and breakthroughs in FSOD research. Moreover, we advocate for the extension of these methods to other tasks such as few-shot instance segmentation and fine-grained object detection, thereby advancing the field of few-shot object detection on all fronts.

References

[1] Y. Cao et al., "Few-shot object detection via association and discrimination," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 16570–16581.

[2] Z. Cao, "Deep learning methods for objective detection," in *Proc. IEEE 2nd Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*, Oct. 2022, pp. 1353–1357.

[3] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "LSTD: A low-shot transfer detector for object detection," in *Proc. AAAI*, vol. 32, 2018, pp. 2836–2843.

[4] Q. Chen, L. Yang, J. Lai, and X. Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4278–4288.

[5] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," 2019, *arXiv:1904.04232*.

[6] M. Cheng, H. Wang, and Y. Long, "Meta-learning-based incremental few-shot object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2158–2169, Apr. 2022.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, Jun. 2014.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[10] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4012–4021.

[11] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3243–3252.

[12] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5311–5320.

[13] J. Han, Y. Ren, J. Ding, K. Yan, and G.-S. Xia, "Few-shot object detection via variational feature aggregation," 2023, *arXiv:2301.13411*.

[14] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10180–10189.

[15] Q. Hu, Y. Gao, and B. Cao, "Curiosity-driven class-incremental learning via adaptive sample selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8660–8673, Dec. 2022.

[16] X. Jiang, Z. Li, M. Tian, J. Liu, S. Yi, and D. Miao, "Few-shot object detection via improved classification features," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5375–5384.

[17] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8419–8428.

[18] L. Karlinsky et al., "RepMet: Representative-based metric learning for classification and few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5192–5201.

[19] P. Kaul, W. Xie, and A. Zisserman, "Label, verify, correct: A simple few shot object detection method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14217–14227.

[20] A. Li and Z. Li, "Transformation invariant few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3093–3101.

[21] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7359–7368.

[22] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 740–755.

[23] L. Liu, B. Ma, Y. Zhang, X. Yi, and H. Li, "AFD-Net: Adaptive fully-dual network for few-shot object detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2549–2557.

[24] Y. Luan, H. Zhao, Z. Yang, and Y. Dai, "MSD: Multi-self-distillation learning via multi-classifiers within deep neural networks," 2019, *arXiv:1911.09418*.

[25] W. L. Luyben, *Distillation Design and Control Using Aspen Simulation*. Hoboken, NJ, USA: Wiley, 2013.

[26] D. Park and J.-M. Lee, "Hierarchical attention network for few-shot object detection via meta-contrastive learning," 2022, *arXiv:2208.07039*.

[27] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled faster R-CNN for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8661–8670.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, vol. 28, 2015, pp. 91–99.

[29] A. A. Rusu et al., "Meta-learning with latent embedding optimization," 2018, *arXiv:1807.05960.*

[30] A. A. Rusu et al., "Meta-learning with latent embedding optimization," in *Proc. ICLR*, 2019.

[31] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[32] K. Song et al., "LightPAFF: A two-stage distillation framework for pre-training and fine-tuning," 2020, *arXiv:2004.12817*.

[33] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7348–7358.

[34] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.

[35] Y. Tang, Z. Cao, Y. Yang, J. Liu, and J. Yu, "Semi-supervised few-shot object detection via adaptive pseudo labeling," *IEEE Trans. Circuits Syst. Video Technol.*, Aug. 2023.

[36] A. Usmanova, F. Portet, P. Lalanda, and G. Vega, "A distillation-based approach integrating continual learning and federated learning for pervasive services," 2021, *arXiv:2109.04197*.

[37] E. Verwimp et al., "Re-examining distillation for continual object detection," 2022, *arXiv:2204.01407*.

[38] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.

[39] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.

[40] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9924–9933.

[41] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Universal-prototype enhancing for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9547–9556.

[42] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 456–472.

[43] S. Wu, W. Pei, D. Mei, F. Chen, J. Tian, and G. Lu, "Multi-faceted distillation of base-novel commonality for few-shot object detection," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 578–594.

[44] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[45] W. Xiong, "CD-FSOD: A benchmark for cross-domain few-shot object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[46] J. Xu, H. Le, and D. Samaras, "Generating features with increased crop-related diversity for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19713–19722.

[47] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9576–9585.

[48] Y. Yang, F. Wei, M. Shi, and G. Li, "Restoring negative information in few-shot object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3521–3532.

[49] D. Zhang, J. Han, G. Guo, and L. Zhao, "Learning object detectors with semi-annotated weak labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3622–3635, Dec. 2019.

[50] G. Zhang, Z. Luo, K. Cui, and S. Lu, "Meta-DETR: Image-level few-shot object detection with inter-class correlation exploitation," 2021, *arXiv:2103.11731*.

[51] J. Zhang, X. Zhang, and Z. Wang, "Task encoding with distribution calibration for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6240–6252, Sep. 2022.

[52] K. Zhang, C. Zhang, S. Li, D. Zeng, and S. Ge, "Student network learning via evolutionary knowledge distillation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2251–2263, Apr. 2022.

[53] L. Zhang, S. Wang, X. Chang, J. Liu, Z. Ge, and Q. Zheng, "Auto-FSL: Searching the attribute consistent network for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1213–1223, Mar. 2022.

[54] Y. Zhao, Q. Ye, W. Wu, C. Shen, and F. Wan, "Generative prompt model for weakly supervised object localization," in *Proc. ICCV*, Oct. 2023, pp. 6351–6361.

[55] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic relation reasoning for shot-stable few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8778–8787.

Han Chen is currently pursuing the master's degree with the School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, China. His research interests include artificial intelligence, computer vision, object detection, image processing, and defect detection.

Qi Wang (Member, IEEE) received the B.E. degree from Northwest Nationality University, Gansu, China, in 2015, the first Ph.D. degree in computer application engineering from the School of Computer Science and Technology, Guangdong University of Technology, Guangdong, China, in 2020, and the second Ph.D. degree in engineering technology from the Faculty of Engineering and Technology, Hasselt University, Hasselt, Belgium, in 2021. He is currently a Special Term Professor with the State Key Laboratory of Public Big Data, Guizhou University. He has authored or coauthored over 20 papers in prestigious conferences and journals in computer vision and multimedia. His current research interests include computer vision, AI security, agricultural vision and text, and vision scope.

Kailin Xie is currently pursuing the master's degree with the School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, China. His research interests include computer vision, object detection, and image segmentation.

Liang Lei received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China. He is currently a Professor with the School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou. His research interests include optical intelligent photoelectric detection, machine equipment, and visual imaging systems.

Matthieu Gaetan Lin received the B.S.E. degree in computer science from ESIEA, Paris, in 2018, and the M.S. degree in computer science from Tsinghua University in 2021. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, under the supervision of Prof. Yong-Jin Liu. His research interests include reinforcement learning and computer vision.

Tian Lv received the B.Eng. degree from Tsinghua University, China, in 2022. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include 3D computer vision, machine learning, and computer graphics.

Yongjin Liu (Senior Member, IEEE) received the B.Eng. degree from Tianjin University, Tianjin, China, in 1998, and the M.Phil. and Ph.D. degrees from The Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is currently a Professor with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computational geometry, computer vision, cognitive computation, and pattern analysis.

Jiebo Luo (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China, Hefei, and the Ph.D. degree in electrical engineering from the University of Rochester, Rochester, NY. His research interests include image processing, pattern recognition, computer vision, medical imaging, and multimedia communication. He was the Co-Chair of the 2007 SPIE International Symposium on Visual Communication and Image Processing. He is on the editorial boards of IEEE TRANSACTIONS ON MULTIMEDIA, *Pattern Recognition*, and the *Journal of Electronic Imaging*.