Fixing Background Misclassification in Few-Shot Object Detection via Product of Experts

Ding Sheng Ong, Yi Liu, Changjing Shang, Guiguang Ding, *Senior Member, IEEE*, Qiang Shen, and Jungong Han, *Senior Member, IEEE*

***Abstract*—Few-shot object detection (FSOD) poses a significant challenge due to the difficulty of learning robust and discriminative object representations under limited supervision. A widely adopted solution is the two-stage fine-tuning framework, wherein knowledge acquired from a large-scale base dataset is transferred to a novel dataset containing only a small number of labeled instances. However, this framework is prone to systematically misclassifying novel objects as background, primarily due to incorrect background label caused by the domain gap between base and novel datasets—an issue exacerbated by the sparse representation of novel categories. In this work, we show that this inherent weakness can be exploited by explicitly redefining the category structure and transferring the representations learned during the base training stage. Building on this insight, we propose a simple yet effective framework grounded in the Product of Experts (PoE) formulation, which estimates the joint distribution over background and novel categories by combining the unnormalized logits from independently trained classifiers. Notably, it does not require modifications of the base model or repetition of the base training phase. Furthermore, we introduce a strategy for identifying additional novel-category instances within the base dataset, which effectively augmenting the training set for fine-tuning. The resulting method is architecture-agnostic, imposes negligible overhead, and integrates seamlessly with existing two-stage fine-tuning pipelines. Extensive experiments on PASCAL VOC and COCO demonstrate that the proposed method yields consistent improvements across different baselines, achieving significant gains over state-of-the-art FSOD approaches.**

***Index Terms*—Few-shot learning, object detection, transfer learning, fine-tuning.**

Received 12 May 2025; revised 27 September 2025; accepted 4 October 2025. Date of publication 17 October 2025; date of current version 9 January 2026. This work was supported in part by the UKRI AIMLAC CDT under Grant EP/S023992/1, in part by the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) through the Welsh Government, in part by the National Natural Science Foundation of China under Grant 62441235 and Grant 62571068, in part by Beijing Natural Science Foundation under Grant L257005, in part by the Qing Lan Project of Jiangsu Universities, and in part by the Major Program of Jiangsu Higher Education Institutions Basic Science (Natural Science) Research under Grant 25KJA520001. Recommended for acceptance by W.-H. Cheng. *(Corresponding author: Jungong Han.)*

Ding Sheng Ong, Changjing Shang, and Qiang Shen are with the Department of Computer Science, Aberystwyth University, SY23 3DB Penglais, U.K.

Yi Liu is with the School of Computer Science and Artificial Intelligence, Aliyun School of Big Data, School of Software, Changzhou University, Changzhou 213000, China.

Guiguang Ding is with the School of Software, Tsinghua University, Beijing 100084, China.

Jungong Han is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: jungonghan77@gmail.com).

Digital Object Identifier 10.1109/TPAMI.2025.3622983

I. Introduction

THE problem of detecting novel object categories given only a few labeled instances imposes a fundamental limitation of current object detection models. Standard object detection frameworks [12], [32], [45] achieve high performance by leveraging large-scale annotated datasets [28], [31], but their reliance on extensive supervision precludes their deployment in scenarios where labeled data is scarce or expensive to acquire.

Several frameworks have been proposed to address this limitation, including Semi-Supervised Object Detection (SSOD) [34], [69], Open-Vocabulary object-Detection (OVD) [14], [30], [71], and Few-Shot Object Detection (FSOD) [25], [55]. Unlike OVD, which benefits from large-scale vision-language pretraining, or SSOD, which leverages unlabeled data, FSOD requires adaptation to novel classes with minimal supervision, typically within 30 annotated examples. The primary challenge lies in learning feature representations that generalize effectively under extreme data scarcity. This problem formulation is particularly relevant in applications such as medical imaging [21], industrial defect detection [9], and rare species identification in biodiversity studies [52], where acquiring extensive labeled data is impractical due to cost, time, or expert availability constraints.

Recent advancements in FSOD have exhibited a clear shift towards transfer learning-based approaches [16], [33], [36], [56], moving away from the meta-learning frameworks that were originally introduced to address the few-shot learning problems [11]. While early FSOD methods predominantly relied on meta-learning [25], [57], [65], subsequent research [55] demonstrated that transfer learning can effectively exploit the rich feature representations learned from large-scale base datasets, enabling efficient adaptation to novel classes without the need for complex meta-learning mechanisms. The simplicity and effectiveness of transfer learning have led to its widespread adoption in FSOD.

Despite its empirical success, transfer learning in FSOD suffers from a fundamental limitation stemming from the partition of categories into base and novel sets. During base training, annotations for novel classes are removed, yet novel objects remain in the images without labels and are therefore treated as background [27], inducing a systematic bias that hinders later recognition. This issue is particularly pronounced in datasets such as MS-COCO [31], where frequent categories (e.g., person, cat, dog) often appear unlabeled, as illustrated in Figs. 2 and 3. Consequently, fine-tuning must overcome the background bias

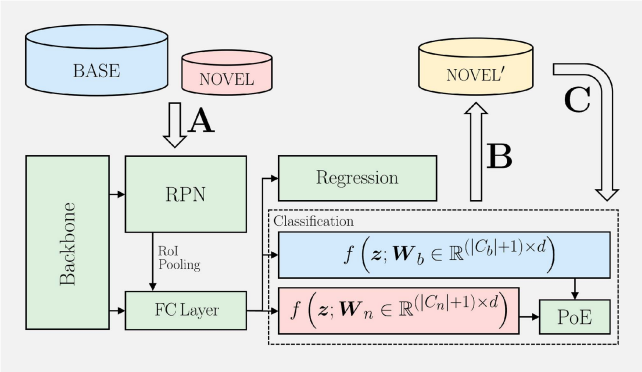

Fig. 1. Overview of the proposed method. (A) We begin by training the object detector using the standard two-stage fine-tuning approach, identical to the baseline, except that we employ a different knowledge transfer mechanism based on the PoE formulation [24]. (B) Using the trained detector, we identify previously unlabeled novel instances within the base dataset and retain the top-k most confident predictions. (C) These additional novel samples are then used to augment the fine-tuning process, enabling more reliable adaptation to novel classes.

under minimal supervision, where features previously assigned to background must be reassigned to novel categories. This contradictory supervision produces conflicting gradients and leads to degraded performance (Fig. 4). The challenge is further exacerbated by the scarcity of novel samples and by restricted fine-tuning strategies. For instance, TFA fine-tunes only the classification head [55], while subsequent methods [27], [42], [50] extend this to the RPN. Although freezing most parameters mitigates overfitting, it limits the model's ability to resolve the semantic inconsistencies introduced by background-label transfer.

Several works have attempted to mitigate this issue indirectly. Qiao et al. [42] introduced a gradient decoupling layer to suppress contradicting gradient signals from the downstream task, preventing the backbone from being adversely influenced during fine-tuning with limited samples. However, as shown in Fig. 4, while this approach reduces the impact of the bias, it does not fully address the challenge of fine-tuning the classifier, which remains susceptible to misclassification. Kaul et al. [27] employed a semi-supervised learning strategy that generates additional labeled samples from unlabeled data using external models such as DINO [68], aiming to mitigate the scarcity of annotated novel examples during feature learning. Similarly, Wang et al. [56] increased novel-category supervision by combining data augmentation with unsupervised saliency detection [40] and CLIP [43]. While effective to some extent, these approaches are far from satisfactorily resolve the underlying issues or rely on external models and auxiliary data, raising concerns about fairness and deviations from the standard FSOD problem formulation.

In contrast to prior work, we propose a training strategy that reframes background misclassification as an opportunity to improve detection. During base training, the background class implicitly absorbs both true background and unlabeled novel instances, causing the classifier to map novel features to background. This assignment is later contradicted during fine-tuning, where the same features are relabeled as novel. To resolve this inconsistency, we redefine the training objectives for the base and novel classifiers independently, resulting in a consistent optimization procedure. As shown in later sections, this formulation naturally leads to an inference rule based on the weighted PoE [24], combining outputs from the background classifier with those of a weaker novel-class classifier trained on limited data.

We further exploit the improved detector obtained after initial fine-tuning to identify unlabeled novel instances in the base dataset. The top-k confident detections are retained as auxiliary supervision for continued training. To enhance robustness against noisy pseudo labels, we employ two mechanisms: confidence-weighted loss scaling and balanced sampling of sourced and ground-truth samples. Despite its simplicity, this procedure yields substantial performance gains without modifying the model architecture or retraining from scratch, thereby avoiding the cost of base-stage reinitialization. As a plug-and-play solution, our approach integrates seamlessly into diverse FSOD baselines, as demonstrated on DeFRCN [42] and MFD [61]. To summarize, our contributions are as follows:

i) We reinterpret the FSOD problem through PoE formulation, in which the outputs of the background and novel classifiers are combined to estimate category likelihoods more accurately, without retraining the baseline model.
ii) We propose a lightweight, plug-and-play training strategy that enhances few-shot adaptation by identifying and incorporating high-confidence novel instances from the base dataset, requiring no additional data, model modifications, or auxiliary networks.
iii) We validate our approach through extensive experiments on standard FSOD benchmarks, including PASCAL VOC and COCO, consistently improving strong baselines and achieving new state-of-the-art performance.

II. Related Work

Object detection is a fundamental problem in computer vision, with existing methods broadly categorized into one-stage [32], [45] and two-stage [12], [13], [23] detectors. While the majority of FSOD research has been developed on two-stage architectures, particularly Faster R-CNN [47], a few studies have explored the feasibility of one-stage detectors such as YOLO [46] and DETR [3]. However, the underlying challenge addressed in this work is independent of detector type, as both categories employ a common mechanism: assigning object categories while implicitly treating undetected regions as background. For instance, DETR's feedforward network (FFN) assigns a special category to queries with no detected objects. In FSOD, this behavior leads to systematic misclassification of novel objects as background during the base training phase, as only a subset of categories is labeled. This bias is a fundamental limitation of the transfer learning framework, yet it remains unaddressed in prior work.

Few-shot learning (FSL) predates FSOD and was originally developed for image-level classification tasks. As a result, early FSOD research was heavily influenced by FSL methodologies [10], [44], [48], [49], [51], [53], [54], particularly through the adoption of meta-learning frameworks [11]. However, FSOD introduces a more complex problem: in addition to classification,

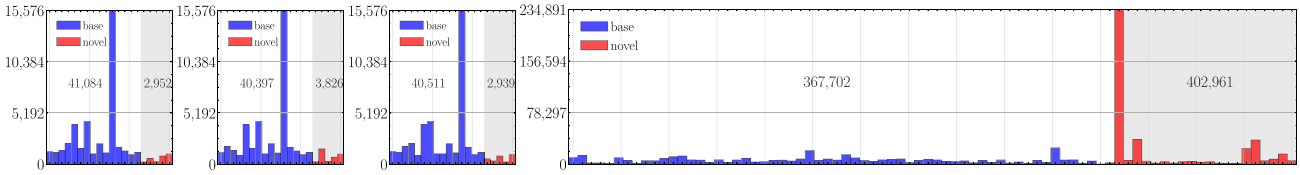

Fig. 2. Presence of unlabeled novel-category instances in base training data. **Blue** indicates annotated *base*-category instances; **red** denotes *novel*-category instances that are present but unlabeled during base training. From left to right: PASCAL VOC Splits 1, 2, 3, and MS COCO. In PASCAL VOC, approximately 10% of objects in the base training set belong to novel categories. In MS COCO, this proportion rises to more than 50%, highlighting the extent of mislabeled samples for novel categories during base training.

Fig. 3. Example of a training image from MS COCO. **Blue** bounding boxes indicate objects from *base* categories, which are annotated during base training. **Red** bounding boxes indicate objects from *novel* categories, which are present but unannotated, and are therefore treated as background by the base classifier.

it requires accurate localization of multiple object instances within a single image. The absence of annotations for certain categories during training further complicates learning, making it difficult to correct misclassifications with only a limited number of labeled instances.

FSOD aims to detect and classify novel objects from limited labeled instances, requiring generalization beyond base categories while preserving localization accuracy. Early approaches relied on meta-learning [11], adapting to novel classes with few annotations [25], [57], [65]. However, these methods involve complex training pipelines, high computational cost, and rigid data organization. More recently, transfer learning approaches [7], [36], [42] have achieved superior performance by leveraging strong representations from large-scale base datasets. As a result, transfer learning has become the dominant paradigm in state-of-the-art FSOD [55]. Unlike meta-learning methods, which model class distributions through comparisons between query and support samples [49], [53], transfer learning methods directly train a classifier for each category. Yet, they suffer from a fundamental issue: systematic misclassification of novel objects as background during base training. While recent works attempt to mitigate this, their solutions remain suboptimal. For example, [61] and [16] augment training with generated samples to reduce classifier bias, while [27] introduces extra data to refine category boundaries, though this approach contradicts the fundamental motivation of FSOD, which assumes minimal supervision.

III. Problem Formulation

This section formalizes the FSOD problem and highlights the key challenges in this setting. We begin with a precise definition (Section III-A) and a review of the widely adopted two-stage fine-tuning paradigm. While effective for transferring knowledge from base to novel categories, this framework exhibits systematic failure modes—-most notably, the misclassification of novel objects as background (Section III-B). To further analyze this phenomenon, we provide qualitative visualizations illustrating how transfer learning strategies contribute to background misclassification, particularly under extreme data scarcity for novel classes (Section III-C).

A. Problem Definition

Following prior work [25], [55], we define the FSOD problem by partitioning a standard object detection dataset, $\mathcal{D} = \{(x_i, y_i)|y_i = \{(b_j, c_j)\}_{j=1}^{M_i}\}$ into two disjoint subsets: a base dataset, $\mathcal{D}_b$ and a novel dataset, $\mathcal{D}_n$. The base dataset contains abundant annotated examples for a set of base categories $\mathcal{C}_b$, while the novel dataset includes a small number of annotated instances for a disjoint set of novel categories $\mathcal{C}_n$, such that $\mathcal{C}_b \cap \mathcal{C}n = \varnothing$. Each image x_i is associated with a set of labeled objects $y_i = \{b_j, c_j\}_{j=1}^{M_i}$, where b_j denotes the bounding box coordinates and c_j indicates the object category.

In the few-shot setting, each dataset provides annotations exclusively for its respective category set. Formally, we define: $\mathcal{D}_b = \{(x_i, y_i)|y_i = \{(b_j, c_j\}_{j=1}^{M_i}, \forall c_j, c_j \in \mathcal{C}_b\}$, and analogously for $\mathcal{D}_n$. Importantly, although images in the original dataset $\mathcal{D}$ may contain both base and novel objects, the partitioning into $\mathcal{D}_b$ and $\mathcal{D}_n$ excludes annotations for categories outside the designated label set. As a result, $\mathcal{D}_b$ may contain unannotated instances of novel categories, which are implicitly labeled as background during base training. This introduces a semantic mismatch: regions containing novel-category objects are treated as background, increasing the risk of misclassification in downstream fine-tuning.

B. Revisiting Two-Stage Fine-Tuning Approach

The two-stage fine-tuning paradigm [55] is a dominant strategy for FSOD. In the first stage, an object detector is trained on the base dataset $\mathcal{D}_b$ to learn category-agnostic representations and classifiers for base classes. In the second stage, the model is fine-tuned jointly on $\mathcal{D}_b$ and a k-shot novel dataset $\mathcal{D}_n$, where each novel category $c \in \mathcal{C}_n$ is represented by exactly k annotated instances: $\forall c \in \mathcal{C}_n, \sum_{(x_i,y_i)\in\mathcal{D}_n} \sum_{(b_j,c_j)\in y_i} \mathbb{1}_c(c_j) = k$. The model parameters from the base stage are reused, while the classifier weights corresponding to novel categories are randomly initialized.

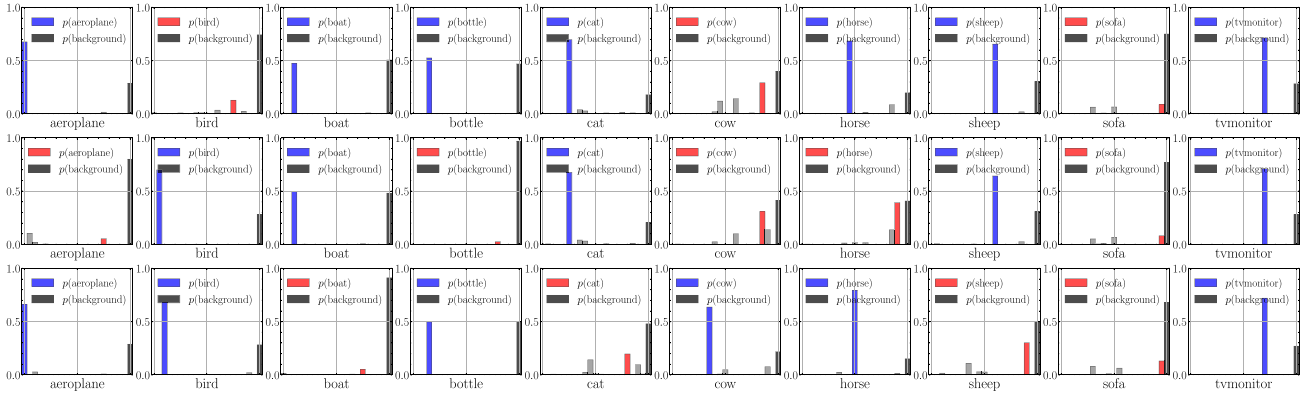

Fig. 4. Predicted probability distirbutions across categories for different dataset splits, using DeFRCN [42] trained on PASCAL VOC. Each row corresponds to a specific data split. The colored bars correspond to the average probability of the category of the instances indicated by the label below the graph. **Blue** represents the *base* categories, and **red** represents *novel* categories. The rightmost bar denotes the average probability of the *background* category. Novel-category instances consistently receive high background probabilities, highlighting systematic misclassification under few-shot settings.

We denote by $f(\boldsymbol{z}; \boldsymbol{W})$ an N-way softmax classifier parameterized by $\boldsymbol{W} \in \mathbb{R}^{N \times d}$, where $\boldsymbol{z} \in \mathbb{R}^d$ is the region-of-interest (RoI) feature vector extracted from the backbone,

$$f(\boldsymbol{z}; \boldsymbol{W})_c = \text{softmax}\left(\boldsymbol{w}_c^\top \boldsymbol{z}\right) = \frac{\exp\left(\boldsymbol{w}_c^\top \boldsymbol{z}\right)}{\sum_{j=1}^{N} \exp\left(\boldsymbol{w}_j^\top \boldsymbol{z}\right)}. \quad (1)$$

Let $f(\boldsymbol{z}; \boldsymbol{W}_b)$ denote the n_b-way softmax classifier used in the base training stage, parameterized by $\boldsymbol{W}_b \in \mathbb{R}^{n_b \times d}$, where the final class corresponds to the background category bg. For notational clarity, we define $n_b = |\mathcal{C}_b| + 1$ and $n_n = |\mathcal{C}_n| + 1$, and we omit the bias term $\boldsymbol{b}_b \in \mathbb{R}^{n_b}$, though it is implicitly included. We propose to reinterpret the background category learned during base training as representing the complement of the base categories, denoted $\mathcal{C}_b'$, which includes all regions not explicitly labeled as base-category instances. This includes both novel-category objects and true background. Formally, since $\mathcal{C}_b \cap \mathcal{C}_n = \varnothing$ and $\text{bg} \notin \mathcal{C}_b$, it follows that: $\mathcal{C}_n \cup \{\text{bg}\} \subseteq \mathcal{C}_b'$. Thus, the base-trained background classifier implicitly models a union of novel categories and background.

C. Background Misclassification

To illustrate the effects of background misclassification, we analyze the output probability distributions of the baseline [42] fine-tuned classifier under the 1-shot setting on PASCAL VOC, where the issue is particularly pronounced due to the extreme scarcity of labeled novel samples. We select ten representative object categories and examine their predicted class probabilities across different dataset splits. As shown in Fig. 4, the predicted distributions for the same category vary significantly depending on whether it is assigned to the base or novel set.

Specifically, when a category is included in the base set, the classifier assigns higher average probability to the correct class than to the background. In contrast, when the same category appears in the novel set, most of its samples are misclassified as background. For example, in splits 1 and 3, the "aeroplane" class belongs to the base set and receives high prediction scores, whereas in split 2, where it is novel, the background probability dominates and the score for the correct class drops sharply.

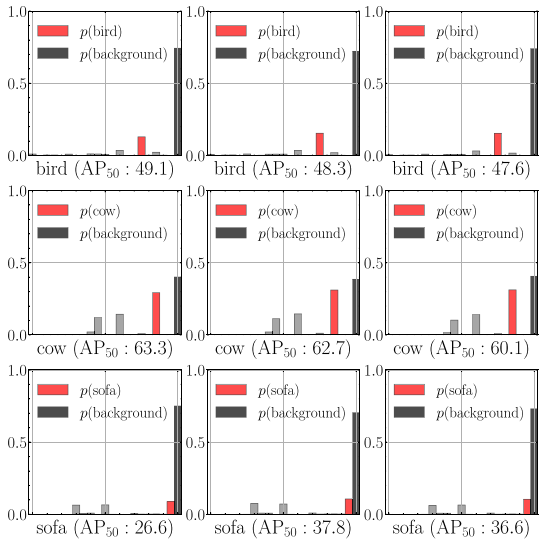

Fig. 5. Predicted probability distributions for selected *novel* categories across different training iterations (DeFRCN [42] on PASCAL VOC Novel Split 1). Each row corresponds to a distinct novel category, and each column shows the predicted probability distribution after training for different numbers of iterations. From left to right: the first column shows the baseline results, while the second and third columns correspond to models trained for $1.5\times$ and $2\times$ the original number of iterations, respectively. The distributions remain largely unchanged across training durations, indicating that extended fine-tuning does not alleviate the background misclassification problem.

This indicates that the classifier fails to separate novel objects from background due to their earlier treatment as unlabeled background during base training. The same trend is observed across multiple categories, underscoring the systematic nature of this misclassification problem.

One might hypothesize that extended training could mitigate this issue by inducing parameter drift sufficient to override the prior bias. In principle, catastrophic forgetting [38] could cause the model to discard the earlier background assignments. In practice, however, this is difficult to achieve in low-shot regimes. As shown in Fig. 5, extending the training schedule by $1.5\times$ or $2\times$ fails to improve performance: the background probability

remains dominant, novel-class scores remain suppressed, and AP50 is unchanged or slightly worse due to overfitting on the support set. These results suggest that background misclassification is not merely a training-duration issue, but a structural limitation of the two-stage fine-tuning paradigm that cannot be addressed by longer training or simple hyperparameter adjustments.

IV. Method

The preceding sections established that background misclassification is a persistent and fundamental issue in few-shot object detection, particularly within the widely adopted two-stage fine-tuning framework [55]. Despite its prevalence in recent methods [7], [15], [16], [36], [42], [56], we have shown that this framework suffers from a semantic inconsistency: categories treated as background during base training may later reappear as labeled novel categories during fine-tuning. In this section, we present a solution to this problem by introducing a new probabilistic framework that more accurately estimates category likelihoods using a PoE [24] formulation, which combines evidence from both the base-trained background classifier and a weak novel classifier (Section IV-A). We leverage the observation that numerous novel-category instances are present in the base dataset by proposing a method for identifying these latent samples and incorporating them into the fine-tuning process, thereby increasing the effective supervision for novel categories without introducing additional data. An overview of the complete method is illustrated in Fig. 1.

A. Product of Experts

As introduced in Section III-B, the conventional two-stage fine-tuning framework transfers all model parameters from the base-trained detector, except for those associated with categories unseen during base training, i.e., the novel categories. These novel-category parameters, including their classifier weights, are typically initialized randomly at the beginning of fine-tuning. This transfer also includes the base-trained background classifier, which we reinterpret as modeling the complement of the base categories, denoted $\mathcal{C}_b'$, and comprising both the true background bg and all novel categories $\mathcal{C}_n$. As previously discussed, this results in a fundamental contradiction: the background classifier, trained to treat novel-category instances as background, is now expected to assign them to explicit novel classes during fine-tuning. This conflict leads to poor adaptation under few-shot supervision.

To avoid the contradiction introduced by transferring background parameters from the base model, we forgo parameter sharing and instead train two classifiers independently. In addition to the n_b-way base classifier, we introduce a separate n_n-way novel classifier that distinguishes novel categories from the true background. Each classifier is trained with a distinct objective. The base classifier is trained under the original supervision regime, where novel-category instances are treated as background. The novel classifier is trained with ground-truth novel labels. Both are optimized using standard negative log-likelihood loss. The combined classification loss is given by:

$$\begin{aligned}\mathcal{L}\text{cls} = & -\sum_{c\in\mathcal{C}_b} \mathbb{1}_c \cdot \log f\left(\boldsymbol{z}; \boldsymbol{W}_b\right)_c \\ & -\alpha \sum_{c\in\mathcal{C}_n\cup\{\text{bg}\}} \log f\left(\boldsymbol{z}; \boldsymbol{W}_b\right)_{n_b} \\ & -(1-\alpha) \sum_{c\in\mathcal{C}_n\cup\{\text{bg}\}} \mathbb{1}_c \cdot \log f\left(\boldsymbol{z}; \boldsymbol{W}_n\right)_c, \end{aligned} \tag{2}$$

where $\alpha \in [0,1]$ controls the contribution of each classifier in learning to distinguish the novel categories and background, which we set to $\alpha = 0.5$ to equal the contributions of both classifiers, and $\mathbb{1}_c$ is an indicator function that equals 1 when the ground-truth label corresponds to class c, and 0 otherwise.

This formulation can be reinterpreted as inference under a weighted product of experts. For base classes, predictions are taken directly from the base classifier. For novel classes and background, the prediction is defined by the sum of log-probabilities from both classifiers:

$$p\left(c \mid \boldsymbol{z}\right) \propto \begin{cases} f\left(\boldsymbol{z}; \boldsymbol{W}_b\right)_c & \text{if } c \in \mathcal{C}_b, \\ f\left(\boldsymbol{z}; \boldsymbol{W}_b\right)_{n_b}^{\alpha} \cdot f\left(\boldsymbol{z}; \boldsymbol{W}_n\right)_c^{1-\alpha} & \text{if } c \in \mathcal{C}_n \cup \{\text{bg}\}. \end{cases} \tag{3}$$

Substituting (3) into the cross-entropy formulation:

$$\mathcal{L}_{\text{cls}} = -\sum_c \mathbb{1}_c \cdot \log p(c \mid \boldsymbol{z}), \tag{4}$$

yields the same loss as in (2). This approach avoids conflicting supervision by isolating the learning signals for base and novel categories, while preserving architectural simplicity and full compatibility with existing FSOD frameworks. We illustrate the conceptual difference between the conventional transfer-learning framework and our proposed method in Fig. 6.

At inference time, final predictions are computed using the weighted PoE formulation in (3). Predictions for base categories are taken directly from the base classifier, while probabilities for novel categories and background are obtained by combining the outputs of the base and novel classifiers. This combination follows a natural probabilistic interpretation, where the base background classifier encodes a negative prior over novel categories, learned during base training when novel instances were treated as background, whereas the novel classifier approximates a likelihood term derived from the few-shot novel set. Their product yields a score proportional to the posterior under this prior–likelihood model, leading to more accurate predictions compared to using the novel classifier alone, which is trained on a limited number of samples.

The weighting parameter $\alpha \in [0,1]$ acts as an uncertainty-calibration factor, log-linearly pooling the base and novel classifiers by scaling their log-probabilities. Smaller α downweights the base background prior and increases reliance on the novel classifier, mitigating the overconfidence of the base classifier on regions containing previously unlabeled novel objects. This weighting provides a simple yet effective mechanism for uncertainty calibration, enabling a smooth trade-off between the base classifier, which reliably detects objects outside the base

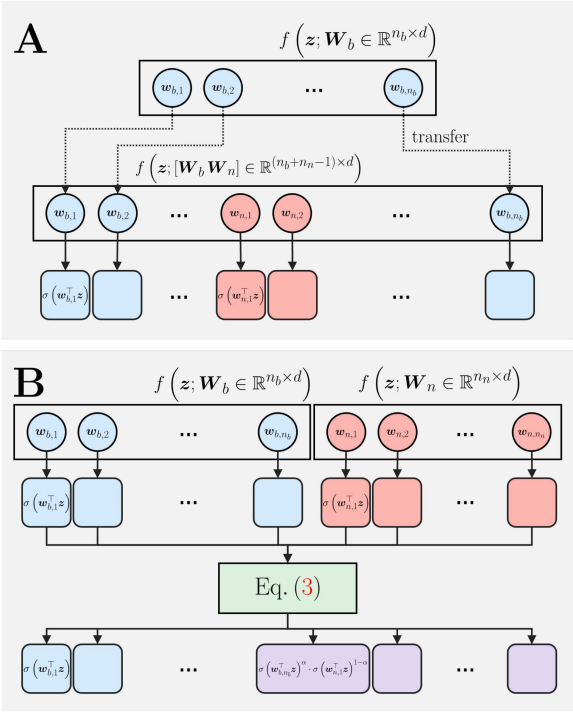

Fig. 6. Conceptual comparison between the conventional two-stage fine-tuning framework [55] and our proposed learning method. (A) In the conventional approach, parameters from the base classifier, including the background class, are transferred, while novel-category weights are randomly initialized. (B) In our framework, a separate novel classifier is initialized independently, and final predictions are obtained via (3), combining outputs from both the base and novel classifiers.

categories, and the novel classifier, which distinguishes among the novel categories and background. A detailed sensitivity analysis of α is presented in Section V-B3.

Our approach is related to ensemble learning in that it combines multiple experts to improve prediction, with the novel classifier acting as a weak expert trained on limited data. However, it differs from conventional averaging, where averaging usually interpolates predictions but does not enforce agreement. By contrast, PoE produces sharper (lower-entropy) predictions and emphasizes consensus between the base and novel classifiers. This property is particularly desirable in FSOD, as it strengthens predictions only when both experts provide consistent evidence, thereby reducing the false negatives commonly observed in previous methods. We empirically compare PoE with simple averaging in Section V-B2 and find that averaging performs worse because its marginalizing effect prevents the novel classifier from reliably distinguishing between novel categories and background under few-shot supervision.

B. Sourcing Additional Samples

To this end, our model enables more accurate classification than conventional transfer learning approaches (see Section V-B). The PoE formulation avoids the contradiction inherent in standard fine-tuning by combining the weak novel classifier with the base classifier, which implicitly models the joint distribution over novel categories and background, $\mathcal{C}_n \cup \{\text{bg}\}$. Although the base classifier cannot explicitly classify novel objects, it has been exposed to them during training and encodes weak but informative features. In contrast, the novel classifier that trained on limited data is prone to overfitting. By combining both sources of evidence, PoE provides a more reliable signal for identifying novel instances. Importantly, this formulation allows us to turn a structural weakness into an advantage without modifying the base training stage.

Since novel-category instances are frequently present but unannotated in the base training set, we leverage this by identifying and recovering such latent samples. These pseudo-labeled novel instances are then used to augment the fine-tuning process, helping to address the limited supervision inherent in few-shot scenarios. However, directly using these samples to train the novel classifier may introduce noise, as some predictions inevitably include false positives. Moreover, due to the scarcity of ground-truth novel annotations, training batches will be dominated by these pseudo-labeled instances, further amplifying their influence. Without proper safeguards, this imbalance can degrade performance rather than improve it. To address these issues, we introduce targeted strategies that regulate the impact of pseudo-labeled samples and ensure that their inclusion contributes positively to the fine-tuning process.

To minimize the risk of introducing noise from falsely identified novel instances, we retain only the top-k most confident predictions per novel category, with $k = 500$ selected based on the ablation results in Section V-B. Nonetheless, some false positives may remain. To mitigate their impact, we incorporate two mechanisms. First, we scale both the classification and localization losses by the predicted confidence score p, allowing high-confidence samples to contribute more to the learning objective, while reducing the influence of uncertain samples:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{rpn}} + p \cdot \mathcal{L}\text{cls} + p \cdot \mathcal{L}_{\text{loc}}, \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss defined in Section IV-A, and $\mathcal{L}_{\text{loc}}$ denotes the localization loss (i.e., Smooth-ℓ1 loss). For ground-truth few-shot samples, we set $p = 1.0$ to ensure they are fully weighted during training. Note that the losses shown are defined per sample, but can be easily extended to a batch of N samples. Specifically, unweighted losses such as $\mathcal{L}_{\text{rpn}}$ are averaged as $\sum_{i=1}^{N} \mathcal{L}_{\text{rpn}}^{(i)}/N$, while confidence-weighted losses are averaged using:

$$\frac{\sum_{i=1}^{N} p^{(i)} \cdot \mathcal{L}_{\text{cls}}^{(i)}}{\sum_{i=1}^{N} p^{(i)}}. \tag{6}$$

Second, we balance the batch composition by sampling both pseudo-labeled and ground-truth novel instances uniformly in each training batch. This guarantees that every batch contains trusted supervision while benefiting from the augmented novel examples. Together, these strategies ensure that false positives have minimal impact on the model while correctly sourced samples reinforce the learning signal, improving generalization to novel categories.

C. Summary

In this section, we introduced a principled framework for addressing the background misclassification problem in few-shot object detection. Our approach is grounded in a PoE formulation that resolves the semantic conflict between base and novel categories by combining the complementary strengths of a robust base classifier and a weak novel classifier. This formulation not only enables consistent probabilistic inference during prediction but also provides a reliable signal for identifying unlabeled novel instances present in the base training set.

To further enhance adaptation, we augment the few-shot support set with high-confidence novel samples discovered from the base dataset. To mitigate the risk of noisy supervision, we introduce two complementary strategies that include: confidence-weighted scaling of classification and localization losses, and balanced sampling of ground-truth and discovered samples during training. These mechanisms allow the model to incorporate additional supervision without modifying the base training stage or increasing architectural complexity. The result is a simple yet effective framework that improves generalization to novel categories while remaining fully compatible with existing FSOD pipelines.

V. Experiments

We evaluate our method on the standard PASCAL VOC [6] and MS COCO [31] benchmarks, following the experimental protocols established in prior work [42], [55], [61]. To assess the effectiveness of each component, we conduct a comprehensive set of ablation studies and qualitative visualizations. These experiments examine the impact of key factors such as the PoE weighting coefficient α, the number of sourced novel samples k, and the contribution of each module to overall detection performance. In addition, we compare our approach against state-of-the-art methods and demonstrate that, when applied to strong FSOD baselines, our framework yields consistent and significant improvements. These results highlight the plug-and-play nature of our method and its practical value in enhancing few-shot object detectors without requiring architectural modifications.

A. Implementation Details

All experiments are conducted using the Detectron2 framework [62] on a single machine equipped with four NVIDIA A100 GPUs. The object detector follows the standard Faster R-CNN architecture [47] with a ResNet-101 backbone [22] as the feature extractor. Training configurations, including learning rate schedules, batch sizes, and other hyperparameters, follow those used in DeFRCN [42] and MFD [61]. Unless otherwise specified, all experiments are run with default settings of $\alpha = 0.5$ and $k = 500$.

Datasets: We evaluate our method on PASCAL VOC [6] and MS COCO [31]. PASCAL VOC consists of 20 object categories, partitioned into 15 base and 5 novel categories across three random splits. All instances from base categories are used during base training, while k-shot examples are randomly sampled per novel class for fine-tuning. Evaluation is performed on the VOC 2007 test set. MS COCO contains 80 categories, of which 20 overlap with VOC and are used as novel classes; the remaining 60 categories form the base set. As with VOC, all base-category instances are used for base training, and k novel-category examples are used in the few-shot stage.

Evaluation Metrics: For PASCAL VOC [6], we report the average precision for novel categories at an IoU threshold of 0.5, denoted as nAP_{50}. For MS COCO [31], we follow the standard evaluation protocol, which includes two widely used reporting strategies. The first is mean average precision (mAP) for novel categories, averaged over IoU thresholds from 0.5 to 0.95, and evaluated across all few-shot settings ($1, 2, 3, 5, 10$, and 30 shots). In addition, several prior works report AP_{50} and AP_{75} exclusively for the 10-shot and 30-shot settings. To ensure fair comparison, we include both types of metrics in our evaluation. Furthermore, we also report results under the generalized few-shot object detection (G-FSOD) setting, a more challenging benchmark that evaluates performance across both base and novel categories. Following prior work [55], we average results over 10 randomly sampled few-shot splits and report AP ($\text{AP}_{0.5:0.95}$), AP_{50}, and AP_{75} for all categories (base + novel), as well as separately for the base and novel subsets. To quantify performance stability in G-FSOD setting, we also report the 95% confidence interval of the mean, computed as $1.96 \cdot s/\sqrt{n}$, where s is the standard deviation and $n = 10$ is the number of repeated runs.

Baselines: To demonstrate the generality of our approach, we integrate it into two strong FSOD baselines: DeFRCN [42] and MFD [61]. Our method requires no changes to the model architecture or parameters as it modifies only the probability computation, as described in Section IV-A. This enables seamless reuse of all pre-trained weights and model files provided by the original authors. Importantly, we do not re-train the base models from scratch and instead use the publicly available pre-trained checkpoints, unless unavailable.

B. Ablation Studies

We conduct ablation studies to evaluate the contribution of each proposed component on the PASCAL VOC Split 1 benchmark using DeFRCN [42] as the base model.

1) Effectiveness of Proposed Components: We first analyze the effectiveness of our two key components, namely the PoE formulation and the sourcing of additional novel samples from the base training set. Throughout this study, we use the default hyperparameters for each component (k and α) as defined in Sections IV-A and IV-B. The impact of varying these hyperparameters is explored in subsequent sections.

As shown in Table I, each component independently improves detection performance across different shot settings. Applying the PoE formulation alone yields an average improvement of 4.1 in nAP_{50} over the baseline, demonstrating that combining signals from the base and novel classifiers leads to better probabilistic estimates. On the other hand, using sourced novel samples without PoE improves performance by 5.6, indicating that augmenting supervision, even with weak labels, is beneficial. When both components are used together, we observe an

TABLE I
ABLATION STUDY ON THE EFFECTIVENESS OF PROPOSED COMPONENTS. WE REPORT nAP_{50} ON PASCAL VOC SPLIT 1 WHEN INDIVIDUALLY APPLYING THE POE FORMULATION (SECTION IV-A) AND ADDITIONAL NOVEL SAMPLE SOURCING (SECTION IV-B), AS WELL AS WHEN COMBINING BOTH.

Method	Sec. 4.1	Sec. 4.2	Novel Split 1					
			1	2	3	5	10	μ
Baseline			57.0	58.6	64.3	67.8	67.0	62.9
PoE only	✓		63.0	66.0	67.3	69.8	68.7	67.0 +4.1
Baseline w/ extra samples		✓	64.6	68.0	68.7	70.5	70.6	68.5 +5.6
PoE w/ extra samples	✓	✓	**66.7**	**69.1**	**70.2**	**70.9**	**70.8**	**69.5** +6.6

overall improvement of 6.6 over using either component alone, confirming their complementary effect.

2) Ablation for Different Mixture Model: From the loss formulation in (2), where the base and novel classifiers are trained independently, we arrive at the PoE inference rule, which corresponds to taking the geometric mean of the probabilities predicted by the two classifiers. This follows naturally from the log-linear structure of the joint objective. To assess whether PoE provides a principled advantage over other fusion strategies, we compare it against two commonly used alternatives: (i) averaging the predicted probabilities, and (ii) a two-stage decision criterion that uses the base classifier's background prediction to gate outputs from the novel classifier. For the averaging method, we modify (3) as:

$$p(c \mid \boldsymbol{z}) \propto \begin{cases} f(\boldsymbol{z}; \boldsymbol{W}_b)_c & \text{if } c \in \mathcal{C}_b \\ \alpha \cdot f(\boldsymbol{z}; \boldsymbol{W}_b)_{n_b} + (1-\alpha) \cdot f(\boldsymbol{z}; \boldsymbol{W}_n)_c & \text{otherwise,} \end{cases} \tag{7}$$

where $\alpha \in [0, 1]$ is a weighting coefficient applied uniformly across classes. For the two-stage decision strategy, the final probability is defined as:

$$p(c \mid \boldsymbol{z}) \propto \begin{cases} f(\boldsymbol{z}; \boldsymbol{W}_b)_c & \text{if } c \in \mathcal{C}_b, \\ f(\boldsymbol{z}; \boldsymbol{W}_n)_c & \text{if } c \in \mathcal{C}_n \cup \{\text{bg}\} \text{ and } f(\boldsymbol{z}; \boldsymbol{W}_b)_{n_b} \geq \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where α is used as a threshold on the base classifier's background probability. This formulation acts as a hard filter that suppresses novel-category predictions when the base classifier is highly confident that a region corresponds to background. However, because hard thresholding is non-differentiable, it cannot be directly incorporated into the cross-entropy loss formulation given in (4) for training. To address this, we employ a differentiable approximation by introducing a sigmoid gating mechanism:

$$p(c \mid \boldsymbol{z}) \propto \begin{cases} f(\boldsymbol{z}; \boldsymbol{W}_b)_c & \text{if } c \in \mathcal{C}_b, \\ \sigma\left(\beta\left(f(\boldsymbol{z}; \boldsymbol{W}_b)_{n_b} - \alpha\right)\right) \cdot f(\boldsymbol{z}; \boldsymbol{W}_n)_c & \text{if } c \in \mathcal{C}_n \cup \{\text{bg}\}, \end{cases} \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function and β is a steepness parameter (set to $\beta = 100$) to closely approximate hard thresholding behavior. In all three mixture strategies, the parameter α governs the influence of the base classifier on novel and background predictions. A higher α increases the dependency on the base classifier, either by assigning it greater weight (in PoE and averaging) or requiring stronger confidence to defer to the novel classifier (in the two-stage criterion).

TABLE II
ABLATION STUDY OF DIFFERENT MIXTURE MODELS. WE REPORT nAP_{50} FOR EACH MIXTURE STRATEGY DEFINED BY (3), (7), AND (9) ACROSS DIFFERENT FEW-SHOT SETTINGS ON PASCAL VOC SPLIT 1.

Model	α	Novel Split 1					
		1	2	3	5	10	μ
Baseline	-	57.0	58.6	64.3	67.8	67.0	62.9
Eq. (3)	0.25	63.0	65.6	66.4	69.7	68.1	66.6 +3.7
	0.50	**63.0**	**66.0**	**67.3**	**69.8**	**68.7**	**67.0** +4.1
	0.75	61.9	65.9	67.7	69.3	67.0	66.4 +3.5
Eq. (7)	0.25	18.2	17.1	18.3	31.4	28.8	22.8 -40.1
	0.50	12.6	12.6	14.3	16.3	20.5	15.3 -47.6
	0.75	12.1	10.0	11.5	10.8	14.3	11.7 -51.2
Eq. (9)	0.25	59.0	61.6	63.3	66.2	65.3	63.1 +0.2
	0.50	58.7	61.7	63.5	66.4	65.7	63.2 +0.3
	0.75	59.3	62.0	63.8	66.7	65.7	63.5 +0.6

The results are summarized in Table II. As predicted, using the PoE formulation (3) achieves the best performance. This is consistent with our earlier analysis, where substituting the PoE rule into the cross-entropy loss yields a training objective that remains aligned with the base training stage, as shown in (5).

Interestingly, the two-stage decision strategy (8) and (9) yields slightly lower performance than PoE, but still improves over the baseline model. In contrast, the weighted averaging method (7) performs significantly worse than the baseline.

The poor performance of the weighted averaging approach can be attributed to its marginalization effect: because it treats the novel and background probabilities as competing terms, high confidence from either classifier can dominate the final prediction. This redundancy weakens the learning signal for the novel classifier, as the base background classifier tends to assign high probability to all novel categories. As a result, the final probability remains high regardless of the novel classifier's output. Consequently, detection accuracy on novel categories degrades significantly, since the base background classifier lacks the capacity to distinguish among different novel classes, and the novel classifier's adaptation is substantially suppressed.

On the other hand, the two-stage decision strategy behaves similarly to conventional training objectives. When substituting (9) into the cross-entropy loss, we obtain:

$$\begin{aligned} \mathcal{L}\text{cls} = & -\sum_{c \in \mathcal{C}_b} \mathbb{1}_c \cdot \log f(\boldsymbol{z}; \boldsymbol{W}_b)_c \\ & -\sum_{c \in \mathcal{C}_n \cup \{\text{bg}\}} \log f(\boldsymbol{z}; \boldsymbol{W}_n)_c \\ & -\sum_{c \in \mathcal{C}_n \cup \{\text{bg}\}} \mathbb{1}_c \cdot \log \sigma\left(\beta\left(f(\boldsymbol{z}; \boldsymbol{W}_b)_{n_b} - \alpha\right)\right), \end{aligned} \tag{10}$$

where the first two terms correspond to standard cross-entropy losses for base and novel classifiers, respectively, and the third term models a gating mechanism. This loss encourages the

TABLE III
ABLATION STUDY ON THE HYPERPARAMETER α, WHICH CONTROLS THE RELATIVE CONTRIBUTIONS OF THE BASE AND NOVEL CLASSIFIERS WHEN LEARNING NOVEL CATEGORIES AND BACKGROUND. AS DESCRIBED IN SECTION IV-A, A HIGHER α INCREASES THE LEARNING RATE ON THE BASE CLASSIFIER DURING TRAINING. WE REPORT nAP_{50} ACROSS DIFFERENT VALUES OF α FOR 1, 2, 3, 5, 10-SHOT SETTINGS ON PASCAL VOC SPLIT 1.

α	Novel Split 1					
	1	2	3	5	10	μ
0 (Baseline)	57.0	58.6	64.3	67.8	67.0	62.9
0.1	61.2	62.6	65.0	68.6	67.5	65.0 +2.1
0.3	62.6	65.7	67.1	69.7	**68.7**	66.8 +3.9
0.5	**63.0**	**66.0**	67.3	69.8	**68.7**	**67.0** +4.1
0.7	62.0	65.3	**67.6**	**70.2**	67.9	66.6 +3.7
0.9	52.4	58.3	61.7	64.7	62.0	60.0 -2.9

TABLE IV
ABLATION STUDY ON THE HYPERPARAMETER k, WHICH CONTROLS THE NUMBER OF TOP-k MOST CONFIDENT SAMPLES SOURCED AS ADDITIONAL SUPPORT FOR FINE-TUNING THE NOVEL CLASSIFIER. WE REPORT nAP_{50} ACROSS DIFFERENT VALUES OF k FOR 1, 2, 3, 5, 10-SHOT SETTINGS ON PASCAL VOC SPLIT 1.

k	Novel Split 1					
	1	2	3	5	10	μ
0 (Baseline)	57.0	58.6	64.3	67.8	67.0	62.9
5	63.4	66.5	66.7	69.3	69.4	67.1 +4.2
10	63.7	66.9	66.9	69.5	68.6	67.1 +4.2
100	**64.8**	67.1	68.8	69.5	70.0	68.0 +5.1
250	64.2	67.6	68.7	70.4	70.3	68.2 +5.3
500	64.6	**68.0**	68.7	**70.5**	**70.6**	**68.5** +5.6
750	64.4	**68.0**	68.8	70.3	70.3	68.4 +5.5
1000	64.1	66.9	**68.9**	70.5	70.4	68.2 +5.3

TABLE V
QUALITY OF SOURCED NOVEL SAMPLES USING DIFFERENT DISCOVERY STRATEGIES ON PASCAL VOC SPLIT 1. WE EVALUATE THE IMPACT OF THE POE FORMULATION ON THE QUALITY OF SOURCED SAMPLES ACROSS DIFFERENT FEW-SHOT SETTINGS (1, 2, 3, 5, AND 10 SHOTS).

Metrics	w/ PoE	Novel Split 1					
		1	2	3	5	10	μ
Accuracy (%)		73.0	77.8	78.6	81.2	82.0	78.5
	✓	**77.7**	**79.1**	**80.0**	**82.4**	**83.6**	**80.6** +2.1
Avg. Confidence, p		0.65	0.60	0.62	0.66	0.67	0.64
	✓	**0.73**	**0.68**	**0.69**	**0.72**	**0.72**	**0.71** +0.07

gating function (thresholding behavior) to "open" for novel and background categories when appropriate.

However, as discussed earlier, the base background classifier is trained to assign high confidence to all novel-category regions. As a result, the sigmoid gating term in the loss function remains high for most inputs, reducing its contribution to the overall optimization objective. Consequently, the two-stage inference rule yields only minimal improvements over the baseline in practice.

3) Ablation for the Hyperparameter α: In this section, we investigate the effect of the weighting parameter α in the PoE formulation. As described in Section IV-A, α controls the balance between the contributions of the base and novel classifiers when modeling novel and background categories. A higher α places greater emphasis on the base classifier during training, effectively slowing the adaptation of the novel classifier. In practice, we set $\alpha = 0.5$ by default, assigning equal weight to both classifiers to ensure balanced learning dynamics. Consistent with previous ablation studies, this experiment excludes the sourcing of additional samples to isolate the effect of α without the influence of auxiliary mechanisms.

As shown in Table III, the optimal value is achieved at $\alpha = 0.5$, which yields the highest average precision across the five few-shot settings. The settings $\alpha = 0.3$ and $\alpha = 0.7$ produce comparable results, slightly lower but still close to the optimal. Interestingly, even when setting $\alpha = 0.1$, we observe improvements over the baseline, suggesting that incorporating signal from the base background classifier remains beneficial for novel-object detection.

When $\alpha = 0$, the contribution of the base classifier is entirely removed, and the model relies solely on the randomly initialized novel classifier, effectively reducing to the baseline two-stage fine-tuning method. On the opposite extreme, setting $\alpha = 0.9$ significantly degrades performance, as the predictions become overly dominated by the base background classifier, which lacks the capacity to distinguish between different novel categories.

4) Ablation for the Hyperparameter k: As described in Section IV-B, we select the top-k most confident samples predicted by the fine-tuned classifier to augment the support set for novel-category fine-tuning. In this experiment, we do not apply the PoE formulation and we use the baseline DeFRCN [42] model to isolate the effects of sample sourcing. All experiments are conducted using balanced sampling strategies and confidence-weighted loss scaling, as previously introduced (Section IV-B).

As shown in the results (Table IV), even sourcing a small number of additional samples significantly improves detection performance—for example, we achieve a $+4.2$ nAP_{50} gain with $k = 5$. Increasing k further continues to improve performance, but beyond $k = 500$, the gains saturate as the fine-tuned classifier struggles to reliably identify additional novel instances. Beyond this point, performance begins to decline slightly, though the drop remains small. This is mitigated by two of our safety mechanisms, i.e., scaling losses by confidence ensures that low-confidence samples have minimal influence on training, and balanced sampling guarantees the inclusion of ground-truth few-shot samples in each batch, helping to suppress noisy gradients.

We further investigate the quality of the sourced novel samples and impact of these sampling strategies in the following section.

5) Quality of Novel Samples: We further analyze the quality of the sourced novel samples. Notably, the PoE formulation enables more accurate sample discovery by leveraging the base background classifier, which has implicitly learned to respond to novel objects during base training. This additional signal improves the reliability of predictions and enhances the quality of the sourced samples used for fine-tuning. To quantify this, we evaluate the sourced samples using a matching criterion: a sample is considered correct if it achieves at least 50% IoU with a ground-truth box and the predicted class label matches. As shown in Table V, samples sourced using the PoE formulation exhibit not only higher accuracy ($+2.1\%$) but also higher

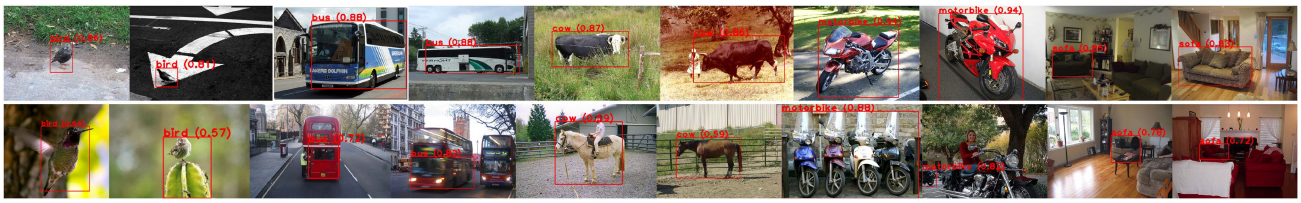

Fig. 7. Examples of additional novel samples sourced from the base dataset on PASCAL VOC Split-1 1-shot using the PoE inference rule described in Section IV-A. The first row shows correctly identified novel samples (verified against ground-truth annotations), while the second row shows incorrectly identified samples.

confidence scores, p $(+0.07)$, which contribute more to the learning objective via the confidence-weighted loss in (5). These results confirm that PoE facilitates more precise discovery of latent novel instances. Together, these findings demonstrate that combining PoE-based inference with novel sample discovery yields substantial improvements in detection performance.

We further illustrate the quality of the sourced samples in Fig. 7, showing examples of correctly and incorrectly identified instances. For each sample, we display the detected bounding box, the predicted label, and the associated confidence score produced by the PoE-based classifier. In the first row, we present samples that were correctly identified and verified against ground-truth annotations. These examples demonstrate high-quality detections with accurate localization and high confidence scores, making them valuable for expanding the novel support set.

Conversely, the second row illustrates incorrectly identified samples. From visual inspection, we identify several sources of error. The majority of false positives arise from imperfect localization, where multiple objects are grouped into a single bounding box or where only a partial object is enclosed, as seen in categories such as bird, bus, and motorbike. In addition, some misclassifications occur due to category confusion, e.g., horses being mistaken for cows, or chairs misclassified as sofas. However, we observe that the confidence scores for these misclassified samples tend to be lower, meaning that the confidence-scaling mechanism naturally suppresses their impact during fine-tuning.

Finally, we note that some apparent errors are attributable to missing or ambiguous ground-truth annotations. For instance, certain motorbike and chair instances were detected but not labeled in the dataset. Such ambiguities suggest that a portion of the "errors" may stem from annotation noise rather than true model failure.

6) Ablation on Sampling Strategies: In this section, we investigate the sourcing of novel samples in greater detail, specifically examining the role and importance of the two safety mechanisms introduced in Section IV-B. We first present results without applying any safety mechanism, followed by ablations where each mechanism is applied individually, and finally results when both mechanisms are combined.

It is important to emphasize that balanced sampling is critical. Without it, the large number of sourced additional samples could dominate the fine-tuning batches, potentially resulting in batches where ground-truth few-shot samples are entirely absent. In such cases, learning would be driven primarily by noisy labels, introducing significant instability where performance may degrade sharply if the sampled labels are inaccurate. Our proposed safety mechanisms are specifically designed to mitigate these risks, enabling more stable and reliable fine-tuning even in the presence of inevitable noise.

TABLE VI
ABLATION STUDY ON THE EFFECTIVENESS OF DIFFERENT SAFETY MECHANISM APPLIED. WE REPORT nAP_{50} ON PASCAL VOC SPLIT 1 WHEN APPLYING CONFIDENCE SCALING AND BALANCED SAMPLING INDIVIDUALLY, AS DESCRIBED IN SECTION IV-B, AND WHEN APPLYING BOTH IN COMBINATION.

Method	Confidence Scaling	Balanced Sampling	Novel Split 1					
			1	2	3	5	10	μ
Baseline			57.0	58.6	64.3	67.8	67.0	62.9
None			58.8	63.9	65.5	69.5	68.4	65.2 +1.3
Confidence Scaling	✓		59.4	64.2	66.1	69.6	68.4	65.6 +1.7
Balanced Sampling		✓	64.5	**68.2**	68.6	69.7	**70.6**	68.3 +5.4
Both	✓	✓	**64.6**	68.0	**68.7**	**70.5**	**70.6**	**68.5** +5.6

In Table VI, we report the effects of different sampling strategies on mitigating the noise from sourced samples. In this experiment, we exclude the PoE formulation introduced in Section IV-A to isolate the effects of sampling, and fix the number of sourced samples to $k = 500$. Without any safety mechanisms, we still observe a slight improvement over the baseline, suggesting that augmenting the support set with additional samples is inherently beneficial. Introducing confidence scaling provides a small further improvement, although its impact remains modest due to the relatively low average confidence of the sourced samples, as shown in Table V. Nonetheless, confidence scaling is necessary to reduce the influence of high-confidence false positives.

Finally, we observe that balanced sampling has the greatest effect in mitigating noise. This confirms our earlier intuition, which is to ensure that each training batch always contains ground-truth samples provides a reliable signal that guides the learning process, even when labeled samples are noisy.

7) One-Stage Detector: In this section, we extend our evaluation to a one-stage detector to further validate the generality of our framework. Specifically, we adopt RetinaNet [32] as a representative one-stage architecture. Since research on transfer learning in FSOD has predominantly focused on two-stage architectures, we construct our own RetinaNet baseline by performing conventional base training followed by novel-class fine-tuning as described in Section III-B. All experiments are conducted under the same settings as outlined in Section V-A, and the results are summarized in Table VII.

One implementation detail is that RetinaNet employs focal loss rather than cross-entropy. Accordingly, we replace the cross-entropy loss in (4) with the focal loss, and estimate the background probability required for the PoE computation

TABLE VII
FEW-SHOT OBJECT DETECTION RESULTS ON PASCAL VOC [6] WITH RETINANET [32]. WE REPORT nAP_{50} ON NOVEL CATEGORIES FOR 1, 2, 3, 5, AND 10 SHOTS, AVERAGED AS μ.

Split	Method	nAP_{50}					
		1	2	3	5	10	μ
1	RetinaNet [32]	20.2	25.6	31.6	36.2	45.7	31.9
	RetinaNet w/ ours	**35.2**	**34.9**	**41.6**	**46.4**	**54.0**	**42.4**
2	RetinaNet [32]	9.7	16.9	25.5	28.3	33.6	22.8
	RetinaNet w/ ours	**12.2**	**20.2**	**33.3**	**36.1**	**40.7**	**28.5**
3	RetinaNet [32]	18.9	20.2	20.1	35.5	37.9	26.5
	RetinaNet w/ ours	**28.5**	**35.9**	**38.1**	**46.9**	**47.8**	**39.4**

in (3) as $\prod_{c\in\mathcal{C}_b}(1-\sigma(\boldsymbol{w}_c^T\boldsymbol{z}))$. From Table VII, we observe consistent improvements across all few-shot settings and data splits, following the same trends reported for two-stage detectors. In particular, our framework yields substantial gains in the low-shot settings, confirming that the core ideas of this work are architecture-agnostic and that explicitly addressing novel objects incorrectly labeled during base training is generally beneficial.

8) Efficiency Analysis: We now consider the computational and memory costs introduced by the proposed framework, relative to a conventional baseline. At a high level, two modifications are introduced: (i) the PoE formulation at the classifier, and (ii) the use of additional novel samples sourced from the base dataset for fine-tuning.

We begin with the PoE formulation. Since all other components of the detector remain unchanged, the additional cost arises solely in the classification layer. In the baseline model, this layer comprises a fully connected transformation followed by a softmax, requiring $\mathcal{O}(Nd)$ FLOPs, where N is the number of categories (e.g. 21 for PASCAL VOC, including background) and d is the feature dimensionality. Under the PoE formulation in (3), we compute two smaller fully connected layers with softmax activations, one over the base classes and one over the novel classes, incurring $\mathcal{O}(n_b d)+\mathcal{O}(n_n d)$ FLOPs. These are followed by elementwise multiplications and exponentiations over the novel categories, adding $\mathcal{O}(n_n)$ FLOPs. Since $N=n_b+n_n-1$, the total cost becomes $\mathcal{O}((N+1)d)+\mathcal{O}(n_n)$, which has the same asymptotic order as $\mathcal{O}(Nd)$. Because $d\gg N$ in typical settings, the matrix multiplication term dominates, so the additional operations constitute only a constant-factor increase. Consequently, the computational overhead is negligible for realistic values of d and N. The memory overhead is likewise minimal, requiring only a single additional d-dimensional weight vector, which is insignificant relative to the parameters of the full detector.

The second source of overhead arises during fine-tuning with additional novel samples. These samples are obtained by running the fine-tuned detector on the base dataset and retaining the top-k highest-confidence detections for each novel category. To maintain the top-k set, we use a heap data structure, resulting in $\mathcal{O}(\log k)$ updates per candidate. This cost is negligible relative to the forward pass of the detector, so the overall time complexity is equivalent to performing inference on $|\mathcal{D}_b|$ images, where $|\mathcal{D}_b|$ is the size of the base dataset. Memory usage scales as $\mathcal{O}(n_n k)$, where n_n is the number of novel categories, as we store only the metadata for the k retained detections per category. The fine-tuning procedure is otherwise unchanged since the batch size, number of iterations, and optimization schedule are identical to the baseline. The only additional computation is a scalar reweighting of the loss by the detection confidence, which contributes a negligible cost.

TABLE VIII
EFFICIENCY ANALYSIS OF THE PROPOSED COMPONENTS. EACH ROW REPORTS THE AVERAGE RUNTIME OF THE CORRESPONDING COMPONENT ACROSS 1, 2, 3, 5, AND 10-SHOT SETTINGS ON PASCAL VOC SPLIT 1, MEASURED ON 4 NVIDIA A100 GPUS. ALL VALUES ARE REPORTED IN SECONDS.

Component	Novel Split 1					
	1	2	3	5	10	Eval. (s/img)
Baseline	295.6 ± 2.5	442.0 ± 5.3	588.4 ± 6.7	742.0 ± 7.0	1188.8 ± 6.3	0.2963
PoE	296.2 ± 2.6	440.8 ± 6.0	592.0 ± 7.5	741.4 ± 6.7	1186.4 ± 10.4	0.2965
Sourcing samples	314.2 ± 4.0	313.4 ± 2.7	314.6 ± 3.2	315.2 ± 2.2	312.6 ± 1.7	-
Fine-tuning	306.4 ± 3.8	455.4 ± 5.3	611.0 ± 4.7	765.6 ± 4.0	1221.4 ± 7.2	0.3060

Table VIII shows the runtime of each component and the baseline on 4 A100 GPUs for PASCAL VOC Split-1, reported over 5 runs. The additional cost introduced by the PoE formulation is negligible, yet it yields a measurable performance improvement (Table I). In contrast, sourcing additional novel samples from the base dataset and performing a second fine-tuning stage incurs a noticeable computational overhead. Nevertheless, as shown in Fig. 5, simply prolonging the fine-tuning schedule does not yield comparable gains, indicating that the proposed training strategy offers a more effective use of computational resources.

9) Quantitative Analysis of Unlabeled Instances: In this section, we systematically analyze the prevalence of unlabeled novel instances in the base dataset and their effect on fine-tuning performance. As shown in Table IX, our method yields the largest performance gains in low-shot settings, where conventional fine-tuning struggles to overcome the systematic background bias introduced by unlabeled novel objects. Notably, the average improvement is highest on Split 2 ($+12.6\%$), compared with Split 1 ($+10.9\%$) and Split 3 ($+9.5\%$). This trend coincides with the number of unlabeled novel instances in the base dataset: Split 2 contains the most, whereas Split 3 contains the fewest. Interestingly, the additional samples sourced in Split 2 exhibit lower average confidence and accuracy, indicating lower quality, yet still lead to the largest improvement. This result suggests that the proposed safety mechanisms effectively mitigate the impact of noisy pseudo-labels during fine-tuning.

C. Comparison to State-of-The-Art

We compare our method against state-of-the-art (SotA) approaches on both the PASCAL VOC and MS COCO benchmarks. Our evaluation includes both FSOD and G-FSOD settings. While the FSOD benchmark is based on single split/shot configurations, the G-FSOD setting evaluates models across 10 randomly sampled few-shot configurations, providing a more robust and reliable comparison.

Among the competing methods, several rely on external data or models. For example, SRR-FSD [72] and Norm-VAE [64] utilize semantic embeddings such as Word2Vec [39], while LVC depends on the DINO [68]. Similarly, SNIDA [56] incorporates

TABLE IX
QUANTITATIVE ANALYSIS OF UNLABELED NOVEL INSTANCES IN THE PASCAL VOC DATASET ACROSS ALL THREE SPLITS AND THEIR IMPACT ON FINE-TUNING. FOR EACH SPLIT, WE REPORT THE NUMBER OF UNLABELED INSTANCES IN THE BASE SET, THE ACCURACY AND CONFIDENCE OF ADDITIONAL NOVEL SAMPLES PREDICTED BY THE FINE-TUNED DETECTOR, AND THE RESULTING NOVEL-CLASS DETECTION PERFORMANCE (nAP_{50}) FOR BOTH THE BASELINE AND OUR METHOD, AND THE RELATIVE IMPROVEMENT.

	Novel Split 1						Novel Split 2						Novel Split 3					
	1	2	3	5	10	μ	1	2	3	5	10	μ	1	2	3	5	10	μ
# Unlabeled Samples	2952						3826						2939					
Accuracy (%)	77.7	79.1	80.0	82.4	83.6	80.6	59.3	64.8	67.2	70.6	73.0	67.0	77.1	77.2	72.5	78.5	79.8	77.0
Avg. Confidence	0.73	0.68	0.69	0.72	0.72	0.71	0.68	0.65	0.67	0.69	0.74	0.69	0.70	0.69	0.69	0.73	0.73	0.71
Baseline (nAP_{50})	57.0	58.6	64.3	67.8	67.0	62.9	35.8	42.7	51.0	54.5	52.9	47.4	52.5	56.6	55.8	60.7	62.5	57.6
Ours (nAP_{50})	66.7	69.1	70.2	70.9	70.8	69.5	45.6	50.1	55.5	56.7	55.8	52.7	61.8	62.5	61.1	64.0	65.4	63.0
Rel. Δ	+17.0%	+17.9%	+9.2%	+4.6%	+5.7%	+10.9%	+27.4%	+17.3%	+8.8%	+4.0%	+5.5%	+12.6%	+17.7%	+10.4%	+9.5%	+5.4%	+4.6%	+9.5%

TABLE X
FEW-SHOT OBJECT DETECTION RESULTS ON PASCAL VOC [6]. WE REPORT MEAN AVERAGE PRECISION ON THE NOVEL CATEGORIES (nAP_{50}). THE BEST AND *SECOND-BEST* RESULTS FOR EACH SETTING ARE HIGHLIGHTED.

Method	Paper	Novel Split 1					Novel Split 2					Novel Split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [4]	AAAI 18	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
FSRW [25]	ICCV 19	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 [57]	ICCV 19	18.9	20.6	20.1	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 [65]	ICCV 19	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 [26]	CVPR 19	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 [66]	NeurIPS 19	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 [55]	ICML 20	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 [60]	ECCV 20	41.7	42.5	51.4	55.2	61.8	24.4	29.3	39.2	39.9	47.8	35.6	41.8	42.3	48.0	49.7
Retentive R-CNN [8]	CVPR 21	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
CME [29]	CVPR 21	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
FSCE [50]	CVPR 21	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 [72]	CVPR 21	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
FADI [2]	NeurIPS 21	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
QA-FewDet [17]	ICCV 21	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
$FSOD^{up}$ [59]	ICCV 21	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	53.5
DeFRCN [42]	ICCV 21	57.0	58.6	64.3	67.8	67.0	35.8	42.7	51.0	54.5	52.9	52.5	56.6	55.8	60.7	62.5
Meta Faster R-CNN [18]	AAAI 22	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
LVC [27]	CVPR 22	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
KFSOD [70]	CVPR 22	44.6	-	54.5	60.9	65.8	37.8	-	43.1	48.1	50.4	34.8	-	44.1	52.7	53.9
FCT [19]	CVPR 22	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
CFA-DeFRCN [15]	CVPR 22	58.2	63.3	65.8	68.9	67.1	37.1	45.5	51.3	55.2	53.8	54.7	57.8	56.9	60.0	63.3
MRSN [37]	ECCV 22	47.6	48.6	57.8	61.9	62.6	31.2	38.3	46.7	47.1	50.6	35.5	30.9	45.6	54.4	57.4
UA-RPN [7]	ECCV 22	40.1	44.2	51.2	62.0	63.0	33.3	33.1	42.3	46.3	52.3	36.1	43.1	43.5	52.0	56.0
KD-TFA++ [41]	ECCV 22	47.0	50.2	52.5	62.1	64.2	29.7	32.9	45.9	48.5	51.1	42.6	46.5	48.8	56.8	57.4
KD-DeFRCN [41]	ECCV 22	58.2	62.5	65.1	68.2	67.4	37.6	45.6	52.0	54.6	53.2	53.8	57.7	58.0	62.4	62.2
MFD [61]	ECCV 22	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
Meta-DETR [67]	TPAMI 22	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
VFA [20]	AAAI 23	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
ICPE [35]	AAAI 23	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	50.9	53.1	55.3	60.6	60.1
NIFF-DeFRCN [16]	CVPR 23	63.5	67.2	68.3	71.1	69.3	37.8	41.9	53.4	56.0	53.5	55.3	60.5	**61.1**	63.7	63.9
Norm-VAE [64]	CVPR 23	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*
σ-ADP [5]	ICCV 23	52.3	55.5	63.1	65.9	66.7	42.7	45.8	48.7	54.8	**56.3**	47.8	51.8	56.8	60.3	62.4
FS-DETR [1]	ICCV 23	45.0	48.5	51.5	52.7	56.1	37.3	41.3	43.4	46.6	49.0	43.8	47.1	50.6	52.1	56.9
FsDetView [63]	TPAMI 23	26.9	35.7	42.3	48.9	57.8	21.2	26.7	30.6	37.7	45.1	24.3	30.4	36.3	41.6	50.1
FPD [58]	AAAI 24	46.5	62.3	65.4	68.2	69.3	32.2	43.6	50.3	52.5	56.1	43.2	53.3	56.7	62.1	64.1
SNIDA-DeFRCN [56]	CVPR 24	59.3	60.8	64.3	65.4	65.6	35.2	40.8	50.2	54.6	50.0	51.6	52.4	55.9	58.5	62.6
SNIDA-MFD [56]	CVPR 24	64.9	67.9	69.7	*71.4*	70.5	42.2	47.8	*54.5*	56.6	54.9	58.1	*61.3*	60.7	*63.6*	**66.0**
DeFRCN w/ ours		*66.7*	**69.1**	*70.2*	70.9	*70.8*	*45.6*	**50.1**	**55.5**	*56.7*	*55.8*	**61.8**	**62.5**	**61.1**	**64.0**	65.4
MFD w/ ours		**67.3**	*68.3*	**70.4**	**71.8**	**70.9**	**45.7**	*49.5*	54.1	**56.8**	*55.8*	*61.1*	60.7	60.8	62.5	65.1

both an unsupervised saliency detector [40] and the CLIP visual-language model [43]. It is also worth noting that all methods use ImageNet-pretrained backbones, while in the case of DeFRCN, this includes the backbone used in its calibration module.

Note that we specifically compare to recent methods that attempt to mitigate the background misclassification issue. LVC [27], for example, identifies additional novel instances using DINO [68] applied to unlabeled images, which is similar to our goal. However, unlike LVC, our approach achieves superior performance by employing a PoE-based inference mechanism to discover novel samples already present in the base training data, without requiring any external model or data. Likewise, SNIDA [56] augments training with synthesized novel-category instances generated using external saliency cues and CLIP embeddings. In contrast, our method sources real novel instances from the base dataset, ensuring higher sample quality while avoiding reliance on external model. This results in more consistent improvements across baselines and shot settings.

1) PASCAL VOC: Table X presents the few-shot detection results across all three standard PASCAL VOC splits. Our method achieves superior performance compared to existing approaches and, more importantly, consistently improves over

TABLE XI
GENERALIZED FEW-SHOT OBJECT DETECTION RESULTS ON PASCAL VOC [6]. FOR EACH METRIC, WE REPORT THE MEAN AND 95% CONFIDENCE INTERVAL, COMPUTED OVER 10 RANDOMLY SAMPLED FEW-SHOT TRAINING SETS.

Split	# shots	Method	Overall #20			Base #15			Novel #5		
			AP	AP_{50}	AP_{75}	bAP	bAP_{50}	bAP_{75}	nAP	nAP_{50}	nAP_{75}
Split 1	1	Meta R-CNN [65]	30.2±0.6	49.4±0.7	32.2±0.9	38.2±0.8	62.6±1.0	40.8±1.1	6.0±0.7	9.9±1.2	6.3±0.8
		TFA w/cos [55]	40.6±0.5	64.5±0.6	44.7±0.6	49.4±0.4	77.6±0.2	54.8±0.5	14.2±1.4	25.3±2.2	14.2±1.8
		DeFRCN [42]	42.8±0.8	67.8±1.4	46.5±0.9	48.9±0.8	75.8±0.7	54.1±1.0	24.4±2.3	43.8±4.3	23.8±2.3
		DeFRCN w/ ours	**48.1±0.4** +5.3	**74.4±0.6** +6.6	**51.8±0.5** +5.3	**54.3±0.1** +5.4	**80.4±0.1** +4.6	**59.9±0.1** +5.8	**29.7±1.5** +5.3	**56.4±2.3** +12.6	**27.6±2.1** +3.8
	2	Meta R-CNN [65]	30.5±0.6	49.4±0.8	32.6±0.7	37.3±0.7	60.7±1.0	40.1±0.9	9.9±0.9	15.6±1.4	10.3±1.0
		TFA w/cos [55]	42.6±0.3	67.1±0.4	47.0±0.4	49.6±0.3	77.3±0.2	55.0±0.4	21.7±1.0	36.4±1.6	22.8±1.3
		DeFRCN [42]	45.1±0.6	71.3±0.8	48.9±0.9	49.2±0.5	75.9±0.7	54.0±0.9	32.8±1.6	57.5±2.5	33.6±2.0
		DeFRCN w/ ours	**50.0±0.1** +4.9	**76.9±0.2** +5.6	**54.2±0.3** +5.3	**54.2±0.1** +5.0	**80.4±0.1** +4.5	**59.9±0.1** +5.9	**37.2±0.6** +4.4	**66.3±0.7** +8.8	**37.3±1.1** +3.7
	3	Meta R-CNN [65]	31.8±0.5	51.4±0.8	34.2±0.6	37.9±0.5	61.3±0.7	40.7±0.6	13.7±1.0	21.6±1.6	14.8±1.1
		TFA w/cos [55]	43.7±0.3	68.5±0.4	48.3±0.4	49.8±0.3	77.3±0.2	55.4±0.4	25.4±0.9	42.1±1.5	27.0±1.2
		DeFRCN [42]	46.1±0.5	72.6±0.5	50.0±0.8	49.6±0.5	76.3±0.6	54.5±0.7	35.5±1.2	61.4±1.7	36.6±1.9
		DeFRCN w/ ours	**50.2±0.1** +4.1	**77.1±0.1** +4.5	**54.4±0.2** +4.4	**54.2±0.1** +4.6	**80.3±0.1** +4.0	**59.6±0.2** +5.1	**38.4±0.6** +2.9	**67.5±0.5** +6.1	**38.6±1.0** +2.0
	5	Meta R-CNN [65]	32.7±0.5	52.5±0.8	35.0±0.6	37.6±0.4	60.6±0.6	40.3±0.5	17.9±1.1	28.0±1.7	19.2±1.3
		TFA w/cos [55]	44.8±0.3	70.1±0.4	49.4±0.4	50.1±0.2	77.4±0.3	55.6±0.3	28.9±0.8	47.9±1.2	30.6±1.0
		DeFRCN [42]	47.0±0.5	73.6±0.5	51.4±0.7	50.0±0.5	76.3±0.4	55.2±0.6	38.3±0.8	65.3±0.9	40.0±1.3
		DeFRCN w/ ours	**50.7±0.1** +3.7	**77.7±0.1** +4.1	**55.2±0.2** +3.8	**54.0±0.1** +4.0	**80.3±0.1** +4.0	**59.7±0.1** +4.5	**40.8±0.4** +2.5	**69.9±0.2** +4.6	**41.5±0.6** +1.5
	10	Meta R-CNN [65]	33.3±0.4	53.8±0.6	35.5±0.4	36.8±0.4	59.8±0.6	39.2±0.4	22.7±0.9	35.6±1.5	24.4±1.0
		TFA w/cos [55]	45.8±0.2	71.3±0.3	50.4±0.3	50.4±0.2	77.5±0.2	55.9±0.3	32.0±0.6	52.8±1.0	33.7±0.7
		DeFRCN [42]	47.5±0.3	74.1±0.5	52.0±0.4	50.0±0.3	76.5±0.3	55.2±0.3	40.2±0.8	67.0±1.4	42.5±1.0
		DeFRCN w/ ours	**50.8±0.1** +3.3	**77.7±0.1** +3.6	**55.1±0.2** +3.1	**53.8±0.1** +3.8	**80.1±0.1** +3.6	**59.1±0.1** +3.9	**41.7±0.5** +1.5	**70.6±0.4** +3.6	**43.1±0.8** +0.6
Split 2	1	Meta R-CNN [65]	30.3±0.5	49.7±0.5	32.3±0.7	38.8±0.6	63.2±0.7	41.6±0.9	5.0±0.6	9.4±1.2	4.5±0.7
		TFA w/cos [55]	36.7±0.6	59.9±0.8	39.3±0.8	45.9±0.7	73.8±0.8	49.8±1.1	9.0±1.2	18.3±2.4	7.8±1.2
		DeFRCN [42]	41.2±0.8	65.2±1.0	44.1±1.0	49.7±0.9	76.5±0.8	54.6±1.2	15.7±2.2	31.5±3.6	12.7±2.4
		DeFRCN w/ ours	**46.1±0.3** +4.9	**71.1±0.5** +5.9	**49.5±0.4** +5.4	**55.0±0.1** +5.3	**81.5±0.1** +5.0	**60.8±0.1** +6.2	**19.1±1.3** +3.4	**39.7±2.0** +8.2	**15.5±1.5** +2.8
	2	Meta R-CNN [65]	30.7±0.5	49.7±0.7	32.9±0.6	38.4±0.5	61.6±0.7	41.4±0.7	7.7±0.8	13.8±1.4	7.4±0.8
		TFA w/cos [55]	39.0±0.4	63.0±0.5	42.1±0.6	47.3±0.4	74.9±0.4	51.9±0.7	14.1±0.9	27.5±1.6	12.7±1.0
		DeFRCN [42]	42.9±0.6	68.0±0.8	45.8±1.0	50.2±0.4	77.1±0.6	55.1±0.8	21.0±1.6	40.9±2.2	18.2±2.4
		DeFRCN w/ ours	**47.2±0.2** +4.3	**72.9±0.3** +4.9	**50.6±0.3** +4.8	**55.0±0.1** +4.8	**81.5±0.1** +4.4	**60.6±0.1** +5.5	**23.9±0.9** +2.9	**46.9±1.2** +6.0	**20.6±1.2** +2.4
	3	Meta R-CNN [65]	31.1±0.3	50.1±0.5	33.2±0.5	38.1±0.4	61.0±0.6	41.2±0.5	9.8±0.9	17.4±1.6	9.4±1.0
		TFA w/cos [55]	40.1±0.3	64.5±0.5	43.3±0.4	48.1±0.3	75.6±0.4	52.9±0.5	16.0±0.8	30.9±1.6	14.4±0.9
		DeFRCN [42]	44.0±0.5	69.2±0.6	47.3±0.7	50.7±0.5	77.1±0.5	55.9±0.7	23.7±1.1	45.6±2.0	21.5±1.4
		DeFRCN w/ ours	**47.6±0.3** +3.6	**73.4±0.3** +4.2	**51.1±0.4** +3.8	**54.9±0.1** +4.2	**81.2±0.1** +4.1	**60.5±0.1** +4.6	**25.6±0.9** +1.9	**50.1±1.2** +4.5	**22.8±1.2** +1.3
	5	Meta R-CNN [65]	31.5±0.3	50.8±0.7	33.6±0.4	37.9±0.4	60.4±0.6	40.8±0.5	12.4±0.9	21.9±1.5	12.1±0.9
		TFA w/cos [55]	40.9±0.4	65.7±0.5	44.1±0.5	48.6±0.4	76.2±0.4	53.3±0.5	17.8±0.8	34.1±1.4	16.2±1.0
		DeFRCN [42]	44.9±0.5	70.6±0.6	48.4±0.7	50.9±0.5	77.4±0.5	56.0±0.7	26.9±1.0	50.1±1.4	25.6±1.4
		DeFRCN w/ ours	**48.2±0.2** +3.3	**74.2±0.2** +3.6	**51.9±0.2** +3.5	**54.8±0.1** +3.9	**81.3±0.1** +3.9	**60.5±0.1** +4.5	**28.1±0.6** +1.2	**53.1±0.9** +3.0	**26.0±0.8** +0.4
	10	Meta R-CNN [65]	32.2±0.3	52.3±0.4	34.1±0.4	37.2±0.3	59.8±0.4	39.9±0.4	17.0±0.8	29.8±1.4	16.7±0.9
		TFA w/cos [55]	42.3±0.3	67.6±0.4	45.7±0.3	49.4±0.2	76.9±0.3	54.5±0.3	20.8±0.6	39.5±1.1	19.2±0.6
		DeFRCN [42]	45.6±0.4	71.3±0.5	48.9±0.4	51.2±0.3	77.5±0.4	55.9±0.3	29.1±0.7	52.8±1.1	28.0±1.1
		DeFRCN w/ ours	**48.4±0.1** +2.8	**74.8±0.2** +3.5	**52.1±0.1** +3.2	**54.5±0.1** +3.3	**80.9±0.1** +3.4	**60.1±0.1** +4.2	**30.2±0.3** +1.1	**56.3±0.6** +3.5	**28.2±0.4** +0.2
Split 3	1	Meta R-CNN [65]	30.8±0.6	49.8±0.8	32.9±0.8	39.6±0.8	63.7±1.0	42.5±0.9	4.5±0.7	8.1±1.3	4.2±0.7
		TFA w/cos [55]	40.1±0.3	63.5±0.6	43.6±0.5	50.2±0.4	78.7±0.2	55.1±0.5	9.6±1.1	17.9±2.0	9.1±1.2
		DeFRCN [42]	42.2±1.1	66.9±2.0	45.5±1.1	49.6±0.6	76.4±0.8	54.6±0.7	20.3±3.9	38.2±6.8	18.2±4.2
		DeFRCN w/ ours	**47.6±0.4** +5.4	**73.6±0.6** +6.7	**51.1±0.5** +5.6	**54.8±0.1** +5.2	**81.1±0.1** +4.7	**60.3±0.2** +5.7	**26.3±1.5** +6.0	**51.2±2.2** +13.0	**23.5±2.0** +5.3
	2	Meta R-CNN [65]	31.3±0.5	50.2±0.9	33.5±0.6	39.1±0.5	62.4±0.9	42.0±0.7	8.0±0.8	13.9±1.4	7.9±0.9
		TFA w/cos [55]	41.8±0.4	65.6±0.6	45.3±0.4	50.7±0.3	78.4±0.2	55.6±0.4	15.1±1.3	27.2±2.1	14.4±1.5
		DeFRCN [42]	44.6±0.6	70.6±0.8	47.8±0.9	50.3±0.3	77.1±0.6	55.1±0.6	27.5±2.0	50.9±2.8	25.9±2.9
		DeFRCN w/ ours	**48.9±0.2** +4.3	**75.6±0.2** +5.0	**52.5±0.4** +4.7	**54.8±0.0** +4.5	**81.2±0.1** +4.1	**60.2±0.1** +5.1	**31.1±0.8** +3.6	**59.0±0.8** +8.1	**29.3±1.3** +3.4
	3	Meta R-CNN [65]	32.1±0.5	51.3±0.8	34.3±0.6	39.1±0.5	62.1±0.7	42.1±0.6	11.1±0.9	19.0±1.5	11.2±1.0
		TFA w/cos [55]	43.1±0.4	67.5±0.5	46.7±0.5	51.1±0.3	78.6±0.2	56.3±0.4	18.9±1.1	34.3±1.7	18.1±1.4
		DeFRCN [42]	45.4±0.5	71.2±0.6	49.3±0.8	50.6±0.4	76.9±0.6	55.8±0.7	30.1±1.4	54.1±2.2	30.0±2.2
		DeFRCN w/ ours	**49.3±0.1** +3.9	**76.1±0.2** +4.9	**53.1±0.3** +3.8	**54.7±0.1** +4.1	**81.2±0.1** +4.3	**60.0±0.2** +4.2	**33.1±0.6** +3.0	**60.9±0.6** +6.8	**32.1±1.1** +2.1
	5	Meta R-CNN [65]	32.4±0.5	51.7±0.8	34.4±0.6	38.5±0.5	61.0±0.7	41.3±0.6	14.0±0.9	23.9±1.7	13.7±0.9
		TFA w/cos [55]	44.1±0.3	69.1±0.4	47.8±0.4	51.3±0.2	78.5±0.3	56.4±0.3	22.8±0.9	40.8±1.4	22.1±1.1
		DeFRCN [42]	46.6±0.3	72.9±0.5	50.4±0.6	51.0±0.3	77.5±0.5	56.1±0.4	33.3±1.2	59.2±1.2	33.5±2.1
		DeFRCN w/ ours	**50.1±0.1** +3.5	**76.9±0.2** +4.0	**54.0±0.2** +3.6	**54.8±0.0** +3.8	**81.2±0.0** +3.7	**60.3±0.1** +4.2	**35.9±0.5** +2.6	**64.0±0.7** +4.8	**35.1±1.0** +1.6
	10	Meta R-CNN [65]	33.1±0.5	53.1±0.7	35.2±0.5	38.0±0.5	60.5±0.7	40.7±0.6	18.4±0.8	31.0±1.2	18.7±1.0
		TFA w/cos [55]	45.0±0.3	70.3±0.4	48.9±0.4	51.6±0.2	78.6±0.2	57.0±0.3	25.4±0.7	45.6±1.1	24.7±1.1
		DeFRCN [42]	47.3±0.3	73.5±0.3	51.3±0.5	51.1±0.3	77.3±0.3	56.2±0.5	35.9±1.0	61.9±1.3	36.6±1.6
		DeFRCN w/ ours	**50.3±0.1** +3.0	**77.1±0.1** +3.6	**54.2±0.2** +2.9	**54.6±0.1** +3.5	**81.0±0.1** +3.7	**59.8±0.1** +3.6	**37.2±0.6** +1.3	**65.4±0.5** +3.5	**37.4±1.0** +0.8

the baseline models—yielding an average gain of $+10.3\%$ over DeFRCN and $+3.8\%$ over MFD across all 15 shot/split combinations. Compared to LVC, which relies on the external DINO model, our method outperforms it by a substantial margin of $+18.2\%$.

When comparing to methods built on the same baseline (e.g., SNIDA and MFD based on DeFRCN), our approach consistently delivers greater improvements and establishes new state-of-the-art results. Specifically, across all 15 shot/split settings, MFD improves the DeFRCN baseline by $+5.3\%$, while SNIDA slightly reduces performance by -1.5%. Focusing on the more challenging 1-shot setting across the three splits, our method achieves an average improvement of nearly $+20\%$, compared to $+11.2\%$ for MFD and only $+0.55\%$ for SNIDA. A similar trend is observed when adopting MFD as the base detector, in which our method improves the MFD baseline by $+4.0\%$ (achieving 61.4 nAP_{50}) compared to a $+2.8\%$ gain by SNIDA (60.7 nAP_{50}). For the 1-shot setting in particular, we achieve an improvement of $+7.7\%$ over the baseline, compared to $+2.2\%$ for SNIDA. Notably, SNIDA relies on additional supervision from external models, including an unsupervised saliency detector [40] and the large-scale visual-language model CLIP [43], neither of

TABLE XII
FEW-SHOT OBJECT DETECTION RESULTS ON MS COCO [31]. WE REPORT THE STANDARD MEAN AVERAGE PRECISION (NAP) FOR ALL SHOT SETTINGS, AS WELL AS NAP_{50} AND NAP_{75} FOR 10-SHOT AND 30-SHOT SETTINGS. THE BEST AND *SECOND-BEST* RESULTS FOR EACH SETTING ARE HIGHLIGHTED.

Method	Paper	1	2	3	5	10			30		
		AP				AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
FSRW [25]	ICCV 19	-	-	-	-	5.6	12.3	4.6	9.1	19.0	7.6
MetaDet [57]	ICCV 19	-	-	-	-	7.1	14.6	6.1	11.3	21.7	8.1
Meta R-CNN [65]	ICCV 19	-	-	-	-	8.7	19.1	6.6	12.4	25.3	10.8
MPSR [60]	ECCV 20	-	-	-	-	9.8	17.9	9.7	14.1	25.4	14.2
TFA w/cos [55]	ICML 20	3.4	4.6	6.6	8.3	10.0	19.1	9.3	13.7	24.9	13.4
FSCE [50]	CVPR 21	-	-	-	-	11.9	-	10.5	16.4	-	16.2
SRR-FSD [72]	CVPR 21	-	-	-	-	11.3	23.0	9.8	14.7	29.2	13.5
FADI [2]	NeurIPS 21	5.7	7.0	8.6	10.1	12.2	-	11.9	16.1	-	15.8
DeFRCN [42]	ICCV 21	6.5	11.8	13.4	15.3	18.6	-	-	22.5	-	-
Meta Faster R-CNN [18]	AAAI 22	5.1	7.6	-	-	12.7	25.7	10.8	16.6	31.8	15.8
KFSOD [70]	CVPR 22	-	-	-	-	18.5	26.3	18.7	-	-	-
FCT [19]	CVPR 22	5.6	7.9	11.1	14.0	17.1	30.2	17.0	21.4	35.5	22.1
CFA-DeFRCN [15]	CVPR 22	-	-	-	15.6	19.1	-	-	23.0	-	-
MFD [61]	ECCV 22	10.8	13.9	15.0	16.4	19.4	-	*20.2*	22.7	-	23.2
KD-TFA++ [41]	ECCV 22	-	-	-	-	12.8	-	11.5	16.6	-	16.1
KD-DeFRCN [41]	ECCV 22	-	-	-	-	18.9	-	17.8	22.6	-	22.6
Meta-DETR [67]	TPAMI 22	7.5	-	13.5	15.4	19.0	30.5	19.7	22.2	35.0	22.8
VFA [20]	AAAI 23	-	-	-	-	16.2	-	-	18.9	-	-
NIFF-DeFRCN [16]	CVPR 23	-	-	-	15.7	19.1	-	-	21.0	-	-
Norm-VAE [64]	CVPR 23	9.5	13.7	14.3	15.9	18.7	-	17.8	22.5	-	22.4
FS-DETR [1]	ICCV 23	7.0	8.9	10.0	10.9	11.3	21.7	11.1	-	-	-
FsDetView [63]	TPAMI 23	-	-	-	-	13.6	28.6	11.3	16.4	32.6	14.7
FPD [58]	AAAI 24	-	-	-	-	16.5	-	-	20.1	-	-
SNIDA-DeFRCN [56]	CVPR 24	10.2	14.5	15.8	16.9	19.1	-	-	23.1	-	-
SNIDA-MFD [56]	CVPR 24	12.0	15.4	*16.4*	*17.8*	*20.7*	-	-	*23.8*	-	-
DeFRCN w/ ours		*12.5*	*15.9*	16.0	17.6	20.2	**36.3**	20.0	23.5	**40.5**	*24.2*
MFD w/ ours		**14.9**	**16.7**	**17.1**	**18.5**	**21.0**	*35.2*	**21.7**	**24.1**	*39.7*	**24.6**

TABLE XIII
GENERALIZED FEW-SHOT OBJECT DETECTION RESULTS ON MS COCO [31]. FOR EACH METRIC, WE REPORT THE MEAN AND 95% CONFIDENCE INTERVAL, COMPUTED OVER 10 RANDOMLY SAMPLED FEW-SHOT TRAINING SETS.

# shots	Method	Overall #80			Base #60			Novel #20		
		AP	AP_{50}	AP_{75}	bAP	bAP_{50}	bAP_{75}	nAP	nAP_{50}	nAP_{75}
1	Meta R-CNN [65]	16.2±0.9	25.8±1.2	17.6±1.0	21.0±1.2	33.3±1.7	23.0±1.4	1.7±0.2	3.3±0.3	1.6±0.2
	TFA w/cos [55]	24.4±0.6	39.8±0.8	26.1±0.8	31.9±0.7	51.8±0.9	34.3±0.9	1.9±0.4	3.8±0.6	1.7±0.5
	DeFRCN [42]	23.9±0.4	36.6±0.7	26.1±0.4	30.2±0.5	45.7±0.8	33.4±0.5	4.8±0.6	9.5±0.9	4.4±0.8
	DeFRCN w/ ours	**31.1±0.1** +7.2	**48.4±0.2** +11.8	**33.1±0.2** +7.0	**38.1±0.1** +7.9	**58.0±0.1** +12.3	**41.2±0.1** +7.8	**10.0±0.4** +5.2	**19.9±0.6** +10.4	**9.0±0.5** +4.6
2	Meta R-CNN [65]	15.8±0.7	25.0±1.1	17.3±0.7	20.0±0.9	31.4±1.5	22.2±1.0	3.1±0.3	6.1±0.6	2.9±0.3
	TFA w/cos [55]	24.9±0.6	40.1±0.9	27.0±0.7	31.9±0.7	50.8±1.1	34.8±0.8	3.9±0.4	7.8±0.7	3.6±0.6
	DeFRCN [42]	25.5±0.5	39.4±0.9	27.8±0.5	31.2±0.4	47.1±0.7	34.5±0.4	8.5±0.9	16.3±1.4	7.8±1.1
	DeFRCN w/ ours	**32.0±0.1** +6.5	**50.1±0.1** +10.7	**33.9±0.1** +6.1	**38.0±0.1** +6.8	**58.0±0.1** +10.9	**41.0±0.1** +6.5	**13.8±0.3** +5.3	**26.5±0.5** +10.2	**12.8±0.5** +5.0
3	Meta R-CNN [65]	15.0±0.7	23.9±1.2	16.4±0.7	18.8±0.9	29.5±1.5	20.7±0.9	3.7±0.4	7.1±0.8	3.5±0.4
	TFA w/cos [55]	25.3±0.6	40.4±1.0	27.6±0.7	32.0±0.7	50.5±1.0	35.1±0.7	5.1±0.6	9.9±0.9	4.8±0.6
	DeFRCN [42]	26.5±0.4	40.9±0.6	28.9±0.3	31.8±0.3	47.9±0.5	35.1±0.3	10.7±0.7	20.0±1.2	10.3±0.8
	DeFRCN w/ ours	**32.4±0.1** +5.9	**50.8±0.1** +9.9	**34.4±0.1** +5.5	**37.9±0.0** +6.1	**57.9±0.1** +10.0	**41.0±0.1** +5.9	**15.6±0.2** +4.9	**29.6±0.4** +9.6	**14.7±0.4** +4.4
5	Meta R-CNN [65]	14.4±0.8	23.0±1.3	15.6±0.8	17.6±0.9	27.8±1.5	19.3±1.0	4.6±0.5	8.7±1.0	4.4±0.6
	TFA w/cos [55]	25.9±0.6	41.2±0.9	28.4±0.6	32.3±0.6	50.5±0.9	35.6±0.6	7.0±0.7	13.3±1.2	6.5±0.7
	DeFRCN [42]	27.7±0.3	42.8±0.5	29.9±0.3	32.4±0.3	48.8±0.4	35.5±0.3	13.5±0.6	24.7±1.1	13.0±0.6
	DeFRCN w/ ours	**32.8±0.0** +5.1	**51.6±0.1** +8.8	**34.8±0.1** +4.9	**37.8±0.1** +5.4	**57.8±0.1** +9.0	**40.7±0.1** +5.2	**17.7±0.2** +4.2	**33.0±0.3** +8.3	**17.1±0.3** +4.1
10	Meta R-CNN [65]	13.4±1.0	21.8±1.7	14.5±0.9	16.1±1.0	25.7±1.8	17.5±1.0	5.5±0.9	0.0±1.6	5.5±0.9
	TFA w/cos [55]	26.6±0.5	42.2±0.8	29.0±0.6	32.4±0.6	50.6±0.9	35.7±0.7	9.1±0.5	17.1±1.1	8.8±0.5
	DeFRCN [42]	29.6±0.2	45.7±0.6	32.1±0.2	33.9±0.2	51.1±0.5	37.2±0.2	16.7±0.5	29.6±1.2	16.7±0.4
	DeFRCN w/ ours	**33.6±0.1** +4.0	**52.7±0.1** +7.0	**35.8±0.1** +3.7	**38.0±0.1** +4.1	**57.9±0.1** +6.8	**41.0±0.1** +3.8	**20.3±0.2** +3.6	**36.8±0.3** +7.2	**20.1±0.2** +3.4
30	Meta R-CNN [65]	13.5±1.0	21.8±1.9	14.5±1.0	15.6±1.0	24.8±1.8	16.9±1.0	7.4±1.1	13.1±2.1	7.4±1.0
	TFA w/cos [55]	28.7±0.4	44.7±0.7	31.5±0.4	34.2±0.4	52.3±0.7	38.0±0.4	12.1±0.4	22.0±0.7	12.0±0.5
	DeFRCN [42]	31.3±0.1	48.5±0.3	33.9±0.1	34.7±0.1	52.4±0.2	38.0±0.2	21.0±0.4	36.7±0.8	21.4±0.4
	DeFRCN w/ ours	**34.1±0.1** +2.8	**53.7±0.1** +5.2	**36.2±0.1** +2.3	**37.6±0.1** +2.9	**57.5±0.1** +5.1	**40.4±0.1** +2.4	**23.7±0.1** +2.7	**42.1±0.2** +5.4	**23.7±0.2** +2.3

which are employed in our approach. In contrast, our method requires no external data or auxiliary pretrained models beyond the original base detector.

We also report results under the G-FSOD setting in Table XI, evaluating performance over 10 randomly sampled few-shot splits. Following prior work, we report both the mean and the 95% confidence interval to reflect result consistency. The G-FSOD setting is more challenging, as it evaluates detection performance across both novel and base categories. As discussed in earlier sections, our PoE formulation explicitly decouples the learning objectives of the base and novel classifiers, avoiding the supervision conflict inherent in conventional transfer-learning frameworks. This not only improves detection for novel categories but also leads to significant gains on base categories. While the relative improvement diminishes as the number of shots increases (e.g., on Split 1, the nAP gain decreases from $+5.3$ at 1-shot to $+1.5$ at 10-shots), the trend is consistent across all splits. Since few prior works report results under G-FSOD, we focus our comparison against DeFRCN. It is also worth noting that the bAP achieved by both the TFA and DeFRCN baselines

remains very similar, further highlighting that the conventional transfer-learning framework inherently alters the learning objectives during fine-tuning, where instances previously labeled as background must be reassigned to novel categories. Our PoE inference rule effectively mitigates this inconsistency, as reflected by the significant improvement in bAP.

2) MS COCO: As previously noted, evaluation metrics in prior COCO-based FSOD studies are inconsistent. Some report COCO-style mean average precision across all shots $(1, 2, 3, 5, 10, 30)$, while others only report AP, AP_{50} and AP_{75} for 10- and 30-shot settings. To ensure comprehensive and fair comparison, we report all of these metrics.

As shown in Table XII, our method consistently improves over the DeFRCN baseline, achieving an average nAP gain of $+20\%$ across all shot settings. Compared with prior methods based on the same backbone, our approach yields substantially larger improvements, in which MFD improves DeFRCN by only $+15.2\%$, and SNIDA achieves a gain of $+13.1\%$. Similarly, when applied to the MFD baseline, our method results in a performance boost of $+10.7\%$, whereas SNIDA, despite leveraging external saliency cues and the CLIP model—achieves only a $+4.5\%$ improvement. Overall, the trend mirrors that observed on the PASCAL VOC benchmark, where our method consistently provides greater improvements over strong baselines without relying on external data or auxiliary models.

We further report results under the G-FSOD setting for MS COCO in Table XIII, using the same evaluation metrics and reporting the 95% confidence interval to assess performance consistency. Our method not only improves detection performance for novel categories but also substantially enhances base-category accuracy. Specifically, we observe that base AP increases by $+17\%$ on average, significantly higher than the gains observed on PASCAL VOC. This further validates our PoE formulation, which effectively mitigates the contradiction in learning objectives between base training and novel fine-tuning stages, particularly in COCO, where a substantial fraction of novel-category objects are unlabeled during base training (see Fig. 4). Additionally, novel AP improves by $+34\%$ on average, with the relative gains gradually decreasing as the number of shots increases (from a $+5.2$ nAP improvement at low shots to $+2.7$ nAP at higher shots).

VI. Conclusion

In this paper, we addressed a fundamental limitation of the two-stage fine-tuning framework for FSOD by redefining the background category as the complement of the defined object classes. Rather than treating novel objects as a source of noise during base training, our approach leverages their implicit presence to improve probability estimation. We redesigned the fine-tuning loss to maintain consistency with the base training objective, naturally deriving a weighted PoE inference rule for estimating the probabilities of novel categories and background. Beyond inference, we further exploited the presence of unlabeled novel instances in the base dataset by proposing a principled strategy to source additional training samples, augmented with safety mechanisms to mitigate the noise introduced during fine-tuning. Extensive experiments on PASCAL VOC and MS COCO benchmarks demonstrate that our method consistently improves performance across strong FSOD baselines, including DeFRCN and MFD. Overall, our findings suggest that biases commonly viewed as limitations in FSOD training can instead be systematically exploited to enhance adaptation under low-data regimes.

References

[1] A. Bulat, R. Guerrero, B. Martinez, and G. Tzimiropoulos, "FS-DETR: Few-shot detection transformer with prompting and without re-training," in *Proc. Int. Conf. Comput. Vis.*, pp. 11793–11802, 2023.

[2] Y. Cao et al., "Few-shot object detection via association and DIscrimination," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 16570–16581.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2020, pp. 213–229.

[4] H. Chen, Y.G. W. Wang, and Y. Qiao, "LSTD: A low-shot transfer detector for object detection," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2018, pp. 2836–2843.

[5] J. Du et al., "Bin he, and jingdong wang. S-adaptive decoupled prototype for few-shot object detection," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 18950–18960.

[6] M. Everingham, L. V. Gool, K. I. ChristopherJ. W. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 98–136, 2010.

[7] Q. Fan, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with model calibration," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 720–739.

[8] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4527–4536.

[9] Z. Fang, X. Wang, H. Li, J. Liu, Q. Hu, and J. Xiao, "FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17481–17490.

[10] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[12] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[14] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–20.

[15] K. Guirguis, A. Hendawy, G. Eskandar, M. Abdelsamad, M. Kayser, and J. Beyerer, "CFA: Constraint-based finetuning approach for generalized few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4038–4048.

[16] K. Guirguis, J. Meier, G. Eskandar, M. Kayser, B. Yang, and J. Beyerer, "NIFF: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24193–24202.

[17] 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. Int. Conf. Comput. Vis.*, 2021, pp. 3263–3272.

[18] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta faster r-CNN: Towards accurate few-shot object detection with attentive feature alignment," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2022, pp. 780–789.

[19] 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.*, 2022, pp. 5311–5320.

[20] J. Han, Y. Ren, J. Ding, K. Yan, and G.-S. Xia, "Few-shot object detection via variational feature aggregation," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2023, pp. 755–763.

[21] N. Hasani et al., "Artificial intelligence in medical imaging and its impact on the rare disease community: Threats, challenges and opportunities," *PET Clin.*, vol. 17, 2022, Art. no. 13.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, 2002.
[25] 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.*, 2019, pp. 8420–8429.
[26] 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.*, 2019, pp. 5197–5206.
[27] 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.*, 2022, pp. 14237–14247.
[28] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.
[29] 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.*, 2021, pp. 7363–7372.
[30] J. Li et al., "Learning background prompts to discover implicit knowledge for open vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16678–16687.
[31] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2014, pp. 740–755.
[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[33] F. Liu et al., "Integrally migrating pre-trained transformer encoder-decoders for visual object detection," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 6825–6834.
[34] L. Liu et al., "MixTeacher: Mining promising labels with mixed scale teacher for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7370–7379.
[35] X. Lu et al., "Breaking immutable: Information-coupled prototype elaboration for few-shot object detection," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2023, pp. 1844–1852.
[36] J. Ma et al., "DiGeo: Discriminative geometry-aware learning for generalized few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3208–3218.
[37] T. Ma et al., "Mutually reinforcing structure with proposal contrastive consistency for few-shot object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 400–416.
[38] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychol. Learn. Motivation*, vol. 24, pp. 109–165, 1989.
[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[40] T. Nguyen et al., "DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 204–214.
[41] W. Pei, S. Wu, D. Mei, F. Chen, J. Tian, and G. Lu, "Few-shot object detection by knowledge distillation using bag-of-visual-words representations," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Heidelberg, 2022, pp. 283–299.
[42] 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.*, 2021, pp. 8681–8690.
[43] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[44] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "Self-supervised knowledge distillation for few-shot learning," in *Proc. 32nd Brit. Mach. Vis. Conf.*, 2021, pp. 1–14.
[45] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[47] S. Ren, K.R. HeGirshick, and J. Sun, "R-CNN faster Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[48] C. Simon, P. Koniusz, and M. Harandi, "Meta-learning for multi-label few-shot classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 346–355.
[49] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4.
[50] 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.*, 2021, pp. 7352–7362.
[51] F. Sung et al., "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.
[52] G. V. Horn et al., "The inaturalist species classification and detection dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769–8778.
[53] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3637–3645.
[54] R.-Q. Wang, X.-Y. Zhang, and C.-L. Liu, "Meta-prototypical learning for domain-agnostic few-shot recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6990–6996, Nov. 2022.
[55] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly simple few-shot object detection," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9919–9928.
[56] Y. Wang, X. Zou, L. Yan, S. Zhong, and J. Zhou, "SNIDA: Unlocking few-shot object detection with non-linear semantic decoupling augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 12544–12553.
[57] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9924–9933.
[58] Z. Wang, B. Yang, H. Yue, and Z. Ma, "Fine-grained prototypes distillation for few-shot object detection," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2024, pp. 5859–5866.
[59] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Universal-prototype enhancing for few-shot object detection," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 9567–9576.
[60] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2020, pp. 456–472.
[61] 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. Eur. Conf. Comput. Vis.*, Cham, 2022, pp. 578–594.
[62] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[63] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3090–3106, Mar. 2023.
[64] 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.*, 2023, pp. 19713–19722.
[65] X. Yan et al., "R-CNN Meta: Towards general solver for instance-level low-shot learning," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9577–9586).
[66] Y. Yang, F. Wei, M. Shi, and G. Li, "Restoring negative information in few-shot object detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. pp. 3521–3532.
[67] G. Zhang, Z. Luo, K. Cui, S. Lu, and E. P. Xing, "Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12832–12843, Nov. 2023.
[68] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–19.
[69] J. Zhang et al., "Semi-DETR: Semi-supervised object detection with detection transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23809–23818.
[70] S. Zhang, L. Wang, N. Murray, and P. Koniusz, "Kernelized few-shot object detection with efficient integral aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19207–19216.
[71] X. Zhao, X. Liu, D. Wang, Y. Gao, and Z. Liu, "Scene-adaptive and region-aware multi-modal prompt for open vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16741–16750.
[72] 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.*, 2021, pp. 8782–8791.

Ding Sheng Ong received the BS degree from the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia, in 2020. He is currently working toward the PhD degree with Aberystwyth University, U.K. His research interests are in few-shot learning in computer vision, particularly object detection.

Yi Liu received the PhD degree from Xidian University, Xi'an, China, in 2019. He is currently a professor with Changzhou University, Changzhou, China. From 2018 to 2019, he was a visiting scholar with Lancaster University, Lancaster, U.K. His research interests include machine learning and computer vision, especially on saliency detection, capsule network, 3-D point cloud, and object detection.

Changjing Shang received the PhD degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1995. She is a University Senior Research Fellow in the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K., and a Fellow of the Learned Society of Wales (the Welsh National Academy). She has published more than 220 peer-reviewed papers and supervised around 30 PhDs/PDRAs. Her research interests include image modelling and analysis, and pattern recognition.

Guiguang Ding (Senior Member, IEEE) is currently a tenured professor in the School of Software with Tsinghua University. He has dedicated himself to promising fields, such as model architecture design and compression, visual semantic recognition and description, transfer learning, and few-shot learning. He has more than 30 papers published in top-tier journals including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Signal Processing Magazine*, and *IEEE Transactions on Image Processing*. Additionally, he has presented more than 70 papers at top-tier international conferences, such as CVPR, NeurIPS, ICML, etc.

Qiang Shen received the PhD degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1990, and an Honorary DSc degree in computational intelligence from Aberystwyth University, U.K., in 2013. He holds the established chair of computer science with Aberystwyth University, is a fellow of the Royal Academy of Engineering, and a recipient of the IEEE Fuzzy Systems Pioneer Award. His research interests include computational intelligence and its applications.

Jungong Han (Senior Member, IEEE) is chair professor with the Department of Automation, Tsinghua University, China. He also holds an Honorary Professorship with Aberystwyth University, U.K. His research interests include computer vision, artificial intelligence, and machine learning. He is a fellow of the International Association of Pattern Recognition, and serves as the associate editor for many prestigious journals, such as *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, etc.