Few-Shot Object Detection via Hierarchical Self-Calibrating and Synergistic Feature Aggregation

1st Yan Wan
School of Information and Intelligent Science
Donghua University
Shanghai, China
winniewan@dhu.edu.cn

2nd Zhiheng Pan
School of Information and Intelligent Science
Donghua University
Shanghai, China
zhiheng925@163.com

***Abstract*—Few-shot object detection (FSOD) addresses data scarcity by learning to detect novel classes from limited examples. Meta-learning has advanced FSOD by enriching class-level representations from scarce support data. However, existing representative methods that improve robustness via distributional feature modeling still face critical bottlenecks: large inter-class bias, weak support–query interaction, and limited generalization capability of the detection head. To address these issues, we propose a Hierarchical Self-Calibrating and Synergistic Feature Aggregation framework. It integrates four calibration stages: 1) centroid-driven feature calibration to suppress intra-class variance; 2) self-attentive Variational Auto-Encoder with Hellinger prior for more reliable class distributions; 3) lightweight bidirectional support–query attention for mutual adaptation; and 4) adaptive prediction logic calibration. To our knowledge, this work is among the first meta-learning-based FSOD frameworks that enable end-to-end collaborative optimization across feature, distribution, interaction, and decision levels. Comprehensive experiments on the PASCAL VOC and MS-COCO benchmarks demonstrate that our model achieves state-of-the-art performance among meta-learning methods, achieving superior performance with only a minor increase in parameters.**

***Index Terms*—few-shot object detection, meta-learning, feature calibration, variational auto-encoder, cross-attention.**

I. Introduction

Few-Shot Object Detection (FSOD) recognizes novel object categories with minimal supervision, reducing annotation dependence. Unlike traditional detectors like Faster R-CNN [1] and YOLO [2] that require large annotated datasets, FSOD uses a two-stage strategy: base-class pre-training and few-shot fine-tuning. Recently, meta-learning serves as a dominant framework, transferring knowledge from base tasks to novel classes with few examples. Despite progress, it still faces challenges including domain gaps, insufficient modeling of intra-class variation, and unstable few-shot optimization.

Early meta-learning methods such as FSRW [3] and Meta R-CNN [4] realign support and query features at the RoI level, yet typically ignore cross-class interactions and intra-class variance. Subsequent approaches like VFA [5] improve robustness by encoding support features as Gaussian distributions and sampling via variational aggregation. Nevertheless, representing categories by averaging few samples remains suboptimal, and modeling distributions via simple networks with KL constraints often proves insufficient.

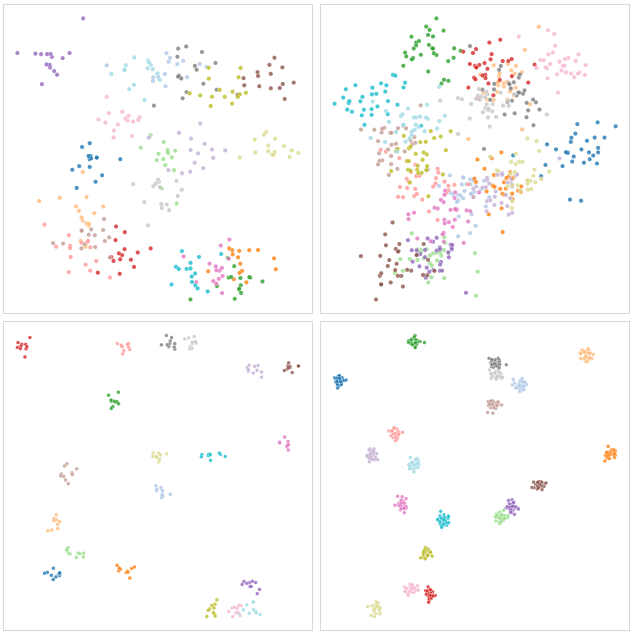

Fig. 1. The t-SNE visualization of support-set RoI features: top row shows 10-shot (left) and 30-shot (right) features; bottom row shows the same features after transformation.

Attention mechanisms aim to leverage support information for prediction calibration in few-shot detection [6]. Yet, in extreme few-shot scenarios (e.g., 1-shot), they often struggle with limited information and become overly reliant on the support prototype, harming generalization [7]. Recent methods like FSAD address this underfitting via cross-attention and adversarial decoupling, boosting 1-shot performance [8]. However, such designs can create excessive support-query coupling as shots increase, leading to unstable performance and underscoring the need for dynamically adaptive interaction. Calibration strategies in FSOD fall into two groups:

feature-level calibration [6], [9], which enhances representation robustness and reduces noise [10], and prediction-level calibration, which rectifies classifier bias from limited samples [11], [12]. Although effective, these are typically studied in isolation. Recently, the Dual Calibration method [13] unifies both within a Faster R-CNN framework, achieving strong results, but it is incompatible with episode-based meta-learning and lacks dynamic support-query interaction.

To overcome these limitations, we propose a Hierarchical Self-Calibrating and Synergistic Feature Aggregation (HSC-SFA) framework. Our principal contributions are fourfold:

1) Proposing a Centroid-Driven Feature Calibration (CFC) module that reduces feature noise and mitigates base-novel distribution overlap;
2) Attentive Variational Auto-Encoder (A-VAE) enhances variational aggregation with self-attention and Hellinger constraints, modeling support features into robust, shot-adaptive distributions;
3) Synergistic Support-Query Attention (SSQA) mechanism enabling mutual adaptive refinement between support and query features, surpassing unidirectional alignment;
4) Adaptive Logic Calibration (ALC) module that fuses cosine similarity with original logits to rebalance detection confidence.

II. Method

The HSC-SFA framework (Fig. 2) achieves hierarchical adjustments to support features and query features within a meta-learning framework. The core components of this framework are elaborated in the following sections.

A. Centroid-Driven Feature Calibration

Meta-learning-based few-shot detectors are prone to noisy backbone features. To mitigate feature noise caused by intra-class variations in few-shot settings, we propose a Centroid-Driven Feature Calibrator (CFC) module. Unlike standard normalization, CFC couples support-informed centroids with local, soft-assigned contraction to suppress noise and tighten class clusters. Concretely, CFC operates in three stages:

Centroid Formation: We maintain a set of global learnable centroids $\mathbf{C}^{\text{global}} = \{\mathbf{c}_k\}_{k=1}^{N} \in \mathbb{R}^{N\times D}$ shared across episodes. For each episode, support prototypes are computed as:

$$\mathbf{p}_c = \frac{1}{|S_c|}\sum_{i\in S_c} f_{\text{roi}}(\mathbf{x}_i) \tag{1}$$

where S_c denotes the support set for class c. Task-specific centroids are formed via concatenation: $\mathbf{C} = \text{Concat}(\mathbf{C}^{\text{global}}, \mathbf{P})$.

Soft Assignment and Momentum Update: The input feature $\mathbf{X} \in \mathbb{R}^{B\times D\times H\times W}$ is flattened to $\tilde{\mathbf{X}} \in \mathbb{R}^{B\times M\times D}$ with $M = H \times W$. For each feature vector $\tilde{\mathbf{X}}_i$, we compute cosine similarity to all centroids:

$$s_{i,j} = \frac{\tilde{\mathbf{X}}_i \cdot \mathbf{c}_j}{\|\tilde{\mathbf{X}}_i\| \cdot \|\mathbf{c}_j\|} \tag{2}$$

A soft assignment weight is computed over the top-K centroids $\mathcal{N}_i$:

$$w_{i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{j'\in\mathcal{N}_i}\exp(s_{i,j'}/\tau)} \tag{3}$$

Global centroids are updated via weighted momentum:

$$\mathbf{c}_k \leftarrow \alpha\mathbf{c}_k + (1-\alpha)\frac{\sum_{i=1}^{M} w_{i,k}\cdot\tilde{\mathbf{X}}_i}{\sum_{i=1}^{M} w_{i,k} + \varepsilon} \tag{4}$$

where $\alpha = 0.1$ is the momentum coefficient.

Contraction Calibration: The weighted displacement vector is computed as:

$$\Delta_i = \sum_{j=1}^{N+K} w_{i,j}(\mathbf{c}_j - \tilde{\mathbf{X}}_i) \tag{5}$$

Then, the calibrated feature is obtained via:

$$\tilde{\mathbf{X}}_i^{\text{calib}} = \text{ReLU}\left(\tilde{\mathbf{X}}_i + \beta\cdot\Delta_i\cdot\exp\left(-\gamma\|\Delta_i\|_2^2/D\right)\right) \tag{6}$$

where $\beta = 1.0$ controls contraction strength and $\gamma = 1.0$ modulates distance sensitivity. The output is reshaped to the original spatial dimensions.

CFC enhances task adaptation via support-informed centroids and reduces feature variance through distance-aware contraction, improving robustness in few-shot detection.

B. Self-Attentive Variational Feature Aggregator

While Variational Auto-Encoder (VAE) has been widely adopted to model support feature distributions [5], standard formulations suffer from two critical limitations in FSOD: neglecting channel-wise dependencies in feature encoding; posterior collapse induced by KL divergence regularization. To address these, we enhance the VAE with two complementary mechanisms.

Channel-Wise Self-Attention: Support RoI features exhibit heterogeneous discriminative power across channels, and redundant channels can degrade distribution modeling—particularly critical for FSOD with limited support samples. We integrate a Squeeze-and-Excitation (SE) block [14] into the encoder to adaptively recalibrate channel responses via modeling inter-channel dependencies. Let $\mathbf{X}$ be the input feature:

$$\mathbf{z} = \text{GAP}(\mathbf{X}), \quad \mathbf{s} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})), \quad \tilde{\mathbf{X}} = \mathbf{s}\odot\mathbf{X} \tag{7}$$

where GAP denotes global average pooling (aggregating global information across channels), δ is ReLU, and σ is Sigmoid. Two fully connected layers ($\mathbf{W}_1$,$\mathbf{W}_2$) capture inter-channel dependencies to generate attention weights, which recalibrate features channel-wise for enhanced discriminability.

Hellinger Distance Regularization: While KL divergence $D_{\text{KL}}(q(z|X)\|p(z))$ regularizes VAE posteriors toward the standard normal prior, its unboundedness causes gradient explosion when scarce support samples yield poorly-estimated posteriors. Moreover, KL's asymmetric penalty precipitates posterior collapse [15], where the posterior degenerates to the prior and forfeits class-specific information. The Hellinger

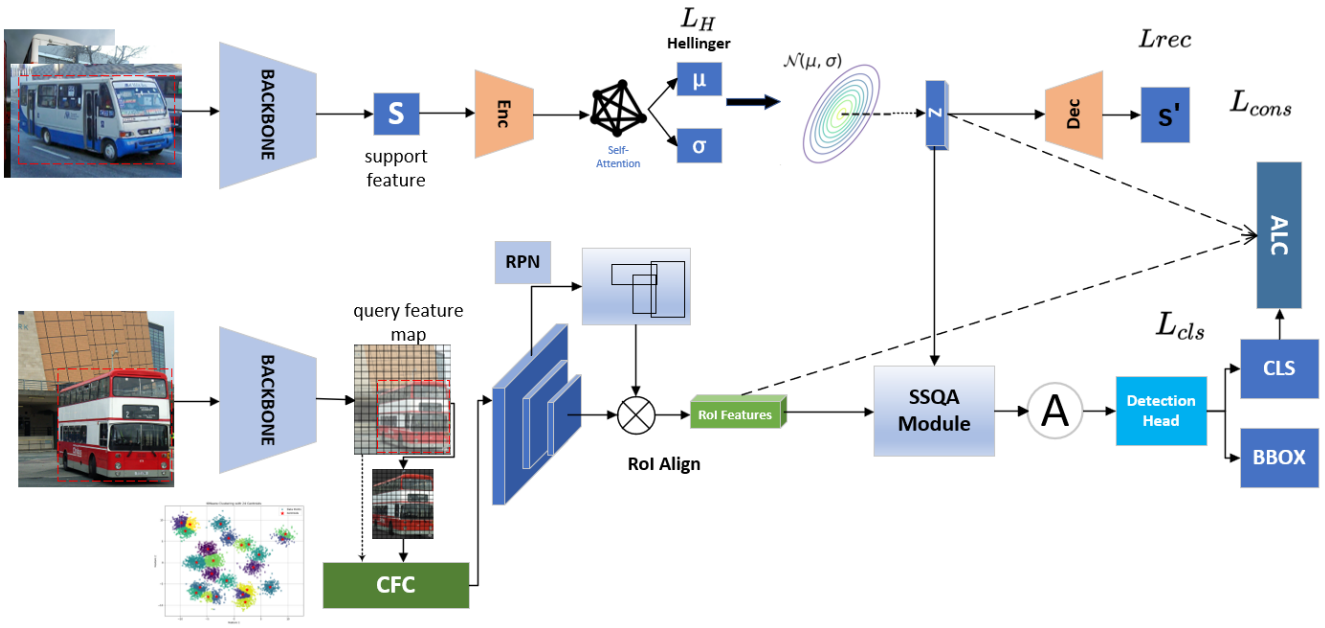

Fig. 2. The overview of our HSC-SFA framework which comprises four principal components. Enc and Dec are the variational feature encoder and decoder.

distance $\mathcal{H}(q,p) \in [0,1]$ provides a bounded, symmetric alternative with smooth gradients [15], stabilizing training while preserving discriminative posterior structure. Given the posterior $q(z|X)$ and prior $p(z) = \mathcal{N}(0, \mathbf{I})$, we compute the distance between sampled latent vectors $\mathbf{z} \sim q(z|X)$ and $\mathbf{z}_{\text{prior}} \sim p(z)$ projected by $\mathbf{P}$:

$$\mathcal{L}_{\text{hell}} = \frac{1}{\sqrt{2}} \left\| \sqrt{\mathbf{P}\mathbf{z}} - \sqrt{\mathbf{z}_{\text{prior}}} \right\|_2 \tag{8}$$

where $\mathbf{P}$ is a learnable projection matrix. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{hell}} \tag{9}$$

with $\lambda = \exp(\omega)$. During base training, ω is learnable; during fine-tuning, it is frozen to prevent overfitting.

A-VAE preserves standard VAE components (encoder, reparameterization, decoder) while generating more discriminative class distributions.

C. Synergistic Support-Query Attention

Building on the robust class distributions modeled by A-VAE, we propose a bidirectional cross-attention mechanism to enable mutual adaptation between query and support features while preventing feature staleness. Unlike unidirectional approaches [20], SSQA realizes symmetric refinement through bidirectional interaction. Crucially, to avoid the quadratic $\mathcal{O}(N^2)$ cost of standard cross-attention (where N is the variable number of support features), SSQA employs a mean-guided, fixed-length memory that reduces the support-side complexity to $\mathcal{O}(1)$ with respect to N.

Query-to-Support Update: The query features $\mathbf{Q}$ attend to the original support features $\mathbf{K}, \mathbf{V}$ to gather class-specific context:

$$\tilde{\mathbf{q}} = \mathbf{q}_{\text{feat}} + \eta \cdot \mathbf{W}_o \left(\text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V} \right), \tag{10}$$

where $\eta = 0.01$ is a residual step size that stabilizes optimization by gating the update magnitude for both directions.

Support-to-Query Update: For lightweight design, instead of support features $\mathbf{k}_s, \mathbf{v}_s$ attending back to the full query set ($\mathcal{O}(M \times N)$), we first condense query information into a single mean vector pair $(\bar{\mathbf{k}}_q, \bar{\mathbf{v}}_q) = \text{Mean}(\mathbf{Q})$, then concatenate with support features to form a fixed-length (length-2) memory:

$$\mathbf{K}_s = [\bar{\mathbf{k}}_q; \mathbf{k}_s], \quad \mathbf{V}_s = [\bar{\mathbf{v}}_q; \mathbf{v}_s] \tag{11}$$

The support features then attend only to this compact memory:

$$\tilde{\mathbf{s}} = \mathbf{s}_{\text{feat}} + \eta \cdot \mathbf{W}_o \left(\text{Softmax} \left(\frac{\mathbf{Q}_s \mathbf{K}_s^\top}{\sqrt{d}} \right) \mathbf{V}_s \right) \tag{12}$$

Since the size of $\mathbf{K}_s$ and $\mathbf{V}_s$ is fixed to 2 regardless of the number of support examples N, this step incurs only $\mathcal{O}(M)$ cost (linear in query size M), and the support-side complexity is $\mathcal{O}(1)$ with respect to N. This makes our SSQA significantly more efficient than standard cross-attention($O(M \times N)$).

The refined features $(\tilde{\mathbf{q}}, \tilde{\mathbf{s}})$ are passed to the subsequent aggregation layer, enabling efficient, bidirectional feature enhancement without a quadratic bottleneck.

D. Adaptive Logic Calibration

To address classifier bias in few-shot settings (which tends to be overconfident for base classes due to abundant training data and underconfident for novel classes due to scarce fine-tuning samples), we design an adaptive logic calibration

TABLE I
FEW-SHOT OBJECT DETECTION PERFORMANCE (nAP_{50}) ON THE PASCAL VOC DATASET. USING RESNET-101 FOLLOWS MOST OF THE PREVIOUS WORKS. BEST SCORES ARE **BOLD** AND SECOND BEST ARE UNDERLINED.

Method / Shot	**Novel Set 1**					**Novel Set 2**					**Novel Set 3**					**Avg.**
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Meta R-CNN [4]	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	31.1
TFA w/ cos [16]	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	39.9
MPSR [17]	41.7	43.1	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	43.4
DeFRCN [11]	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4	51.9
Meta FR-CNN [7]	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	50.5
ICPE [18]	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.5	52.5	50.9	53.1	55.3	60.6	60.1	54.95
FSAD [8]	54.7	61.2	62.5	65.1	65.8	35.3	47.1	50.3	52.2	52.7	46.6	55.0	55.8	60.2	60.3	54.99
FPD [19]	48.1	62.2	64.0	67.6	68.4	29.8	43.2	47.7	52.0	53.9	44.9	53.8	**58.1**	61.6	**62.9**	54.55
VFA [5]	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	56.1
HSC-SFA (Ours)	**58.6**	**67.1**	**68.0**	**68.6**	**69.1**	**44.2**	**51.2**	**53.2**	**53.8**	**55.0**	**51.4**	**56.0**	57.6	**62.0**	62.1	**60.35**

module to rebalance detection confidence by fusing two complementary signals: original logits and cosine similarity.

Feature Projection: Query and prototype features are projected via a shared linear layer:

$$\mathbf{q} = \mathbf{W}\mathbf{f}_q, \quad \mathbf{p} = \mathbf{W}\mathbf{f}_p \tag{13}$$

Cosine Similarity and Scaling: The cosine similarity and its magnitude scale are computed as:

$$\pi = \frac{\mathbf{q}^\top \mathbf{p}}{\|\mathbf{q}\| \cdot \|\mathbf{p}\|}, \quad \tau_{\text{cls}} = \|\mathbf{z}_{\text{raw}}\|_2, \quad \tau_{\text{cos}} = |\pi| \tag{14}$$

where $\mathbf{z}_{\text{raw}} \in \mathbb{R}^{C+1}$ is the vector of original logits (including background).

Learnable Fusion: Foreground logits are recalibrated through an adaptive fusion:

$$\hat{z}_{\text{fg}} = w_{\text{raw}} \cdot z_{\text{raw}} \cdot \tau_{\text{cos}} + w_{\text{cos}} \cdot \pi \cdot \tau_{\text{cls}} \tag{15}$$

where w_{raw} and w_{cos} are obtained by normalizing two independent learnable parameters $\omega_{\text{raw}}, \omega_{\text{cos}} \in \mathbb{R}$:

$$(w_{\text{raw}}, w_{\text{cos}}) = \text{softmax}([\omega_{\text{raw}}, \omega_{\text{cos}}]) \tag{16}$$

Background scores are adjusted by transferring the magnitude of the original prediction:

$$\hat{z}_{\text{bg}} = z_{\text{raw,bg}} + \tau_{\text{cls}} \tag{17}$$

The fusion weights are frozen during fine-tuning to ensure stability.

ALC fine-tunes overconfident logits in the detection head, thereby compensating for underfitting or overfitting issues in detection models while avoiding additional data dependency and branches.

III. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate the proposed method on two widely-used FSOD benchmarks: PASCAL VOC [21] and MS-COCO [22]. The data splits and evaluation protocols follow the established settings from TFA [16] for fair comparison.

Dataset: VOC contains 20 object categories. Following the standard protocol [4], we divide the categories into three splits, each with 15 base classes and 5 novel classes. We evaluate with $K \in \{1, 2, 3, 5, 10\}$ shots per novel class. MS-COCO includes 80 object categories. We adopt the 60/20 split introduced by [16]:60 base classes and 20 novel classes, and evaluate at $K = 10$ and 30 shots.

Evaluation Metrics: On VOC we report *bAP* (base) and *nAP* (novel) at IoU $= 0.5$. On COCO we follow the official IoU $= 0.5 : 0.95$ metric and only report the mean AP of novel classes(*nAP*).

TABLE II
COMPARISON WITH THE META-LEARNING METHODS ON MS-COCO DATASET(NAP).

Method	**10-shot**	**30-shot**
FSDetView [23]	12.5	14.7
Meta FR-CNN [7]	12.7	16.6
DCNet [24]	12.8	18.6
CME [25]	15.1	16.9
FSAD [8]	16.1	18.7
FPD [19]	15.9	19.3
VFA [5]	16.2	18.9
HSC-SFA (Ours)	**18.8**	**22.2**

B. Implementation Details

We implement our framework using MMDetection [26] with a ResNet-101 [27] backbone pre-trained on ImageNet-1K [28], conducting all experiments on four NVIDIA V100 GPUs. For base training we use SGD with learning rate 0.01 and weight decay 1×10^{-4}, reducing to 0.001 for novel class fine-tuning while freezing critical parameters (Hellinger weight w and ALC fusion weights) for stability. The centroid momentum α in CFC is set to 0.1, and centroid number 24 achieves the best balance between representation capacity and computational efficiency. Our training schedule follows Meta R-CNN [4]: 400, 800, 1200, 1600, 2000 iterations for $K = \{1, 2, 3, 5, 10\}$-shot VOC respectively; and 10,000/20,000 iterations for 10/30-shot COCO settings. All remaining training details are kept identical to the original Meta R-CNN [4] configuration to eliminate confounding factors.

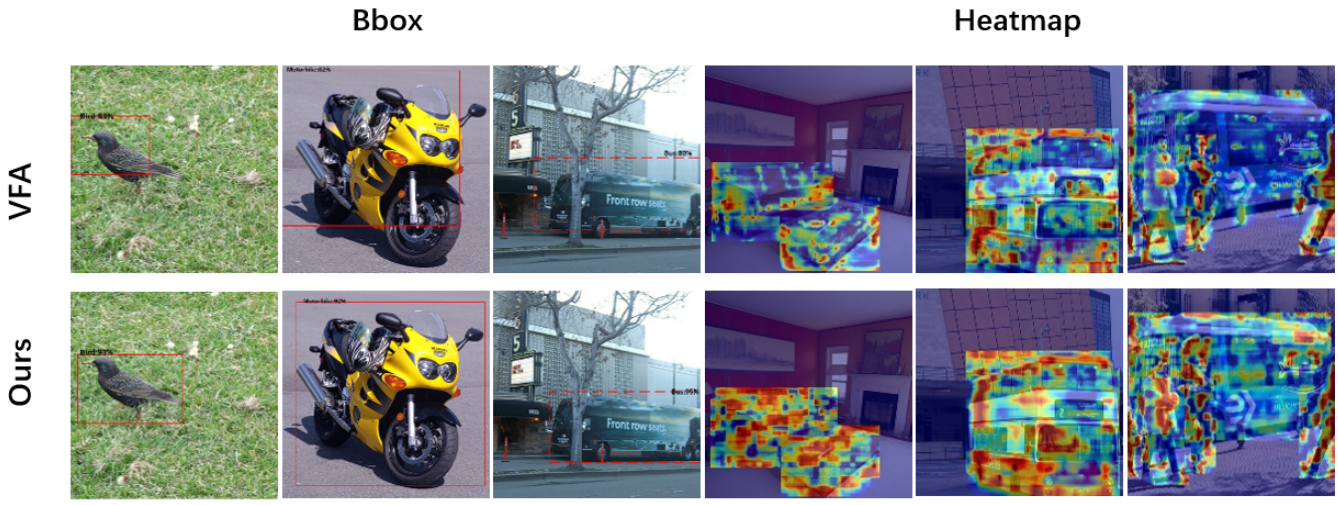

Fig. 3. Comparison of bbox and heatmap visualization results between the VFA method and our approach.

C. Parametric Statistics

The added modules contain approximately 4M learnable parameters, accounting for less than 8% of the base Meta R-CNN [4]. In terms of computational complexity, the additional computational cost is 0.727 GFLOPs, only 0.86% of the base model. These results demonstrate that the proposed modules achieve feature calibration, logical fusion, and attention enhancement with only slight increases in parameter count and computational overhead. They incur minimal storage and inference costs, offering excellent practicality for resource-constrained few-shot detection scenarios.

D. ExperimentalResults

Table I presents the comprehensive comparison on PASCAL VOC. Among mainstream FSOD methods based on meta-learning (excluding LLM-augmented methods [29]), our method achieves state-of-the-art performance across all shot settings and novel splits. Specifically, while the baseline achieves *bAP* of approximately 79%, our HSC-SFA framework not only improves this by around 1% but also delivers remarkable performance on novel classes under few-shot scenarios: HSC-SFA outperforms the strong VFA baseline by 0.9–3.3% on Novel Set 1 and shows significant improvements of up to 5.0% on the challenging Novel Set 2. The average performance across all three splits reaches 60.35% nAP_{50}, surpassing VFA by 4.25%.

More extensively on MS-COCO, our method obtains competitive performance of 18.8% and 22.2% nAP in 10-shot and 30-shot settings respectively (Table II), demonstrating strong generalization capability in complex detection scenarios. To isolate the contribution of our feature–aggregation strategy, we deliberately omit advanced fine-tuning techniques such as the gradient-decoupling layer adopted in DeFRCN [11].

The t-SNE plot of support-set RoI features in Fig. 1 reveals clear inter-class separation and tight intra-class clustering, indicating superior feature discriminability. Correspondingly, the bounding-box detection and prediction heatmap results in Fig. 3 confirm that these well-structured features translate into more precise localization and accurate recognition, even in few-shot settings.

TABLE III
ABLATION STUDY OF DIFFERENT COMPONENTS ON PASCAL VOC (AVERAGE OVER THREE SPLITS).

Components	**nAP_{50}**	**Δ**
VFA Baseline	56.10	-
+ A-VAE	57.60	+1.50
+ A-VAE + CFC	58.85	+2.75
+ A-VAE + SSQA	57.85	+1.75
+ A-VAE + CFC + SSQA	59.52	+3.42
+ A-VAE + CFC + ALC	59.30	+3.20
+ A-VAE + CFC + SSQA + ALC	**60.35**	**+4.25**

E. Ablation Study

To quantitatively evaluate the independent contribution of each component and their synergistic effects, we conduct systematic ablation studies on PASCAL VOC, with results summarized in Table III. Starting from the VFA baseline, replacing the variational feature aggregator with A-VAE improves the baseline by 1.50% nAP_{50}, attributed to channel-wise attention and Hellinger regularization that enhance feature discrimination in higher-shot scenarios. Incorporating CFC further boosts performance to 58.85% (+2.75% cumulative), effectively suppressing feature noise and sharpening class boundaries, particularly when sufficient support samples are available.

Experimental results demonstrate that SSQA and ALC are functionally complementary: adding SSQA to A-VAE yields 57.85% (+1.75%), while introducing ALC to the A-VAE + CFC combination achieves 59.30% (+3.20%). The complete configuration (A-VAE + CFC + SSQA + ALC) achieves

optimal performance at 60.35% nAP_{50}, delivering a substantial +4.25% improvement over the baseline with minimal computational overhead.

This confirms that the hierarchical integration of all core modules enables synergistic optimization of feature representation, distribution modeling, support-query interaction, and decision-making logic, thereby maximizing the model's few-shot detection performance across all shot settings.

IV. Conclusion

This paper proposes the HSC-SFA framework to address core challenges in FSOD: insufficient feature discriminability, weak support-query interaction, and unstable prediction confidence. Our method integrates four novel complementary components for feature noise suppression, robust class distribution modeling, mutual support-query feature adaptation, and prediction confidence rebalancing, forming a hierarchical self-calibration mechanism that addresses FSOD bottlenecks across multiple dimensions. Extensive experiments on standard FSOD benchmarks show that HSC-SFA achieves state-of-the-art performance, outperforming existing advanced meta-learning methods. Notably, this framework provides a scalable meta-learning solution with end-to-end collaborative optimization of multi-level calibration, offering a new design perspective for calibration-based FSOD and a foundation for more stable few-shot learning systems. Future work will explore integrating HSC-SFA with vision-language pre-training to enable zero-shot transfer to unseen semantic categories.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8419–8428.

[4] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9576–9585.

[5] J. Han, Y. Ren, J. Ding, K. Yan, and G.-S. Xia, "Few-shot object detection via variational feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1. AAAI Press, 2023, pp. 755–763.

[6] T. I. Chen, Y. C. Liu, H. T. Su, Y. C. Chang, Y. H. Lin, J. F. Yeh, W. C. Chen, and W. H. Hsu, "Dual-awareness attention for few-shot object detection," *Multimedia, IEEE Trans. on (T-MM)*, vol. 25, no. 000, p. 11, 2023.

[7] 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 *AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233289453

[8] Y. Kou, K. Wu, C. Huang, H. Chen, W. Du, and H. Liu, "Fsad:few-shot object detection via aggregation and disentanglement," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[9] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "Accurate few-shot object detection with support-query mutual guidance and hybrid loss," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 419–14 427.

[10] S. Zhang, L. Wang, N. Murray, and P. Koniusz, "Kernelized few-shot object detection with efficient integral aggregation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 185–19 194.

[11] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8661–8670.

[12] H. Chen, Q. Wang, K. Xie, L. Lei, M. G. Lin, T. Lv, Y. Liu, and J. Luo, "Sd-fsod: Self-distillation paradigm via distribution calibration for few-shot object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5963–5976, 2024.

[13] D. S. Ong, Y. Liu, and J. Han, "Towards few-shot object detection through dual calibration," *IEEE Transactions on Intelligent Vehicles*, pp. 1–14, 2024.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[15] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, and Z. Tu, "Guided variational autoencoder for disentanglement learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7917–7926.

[16] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *37th International Conference on Machine Learning: ICML 2020, Online, 13-18 July 2020, Part 13 of 15*, 2021.

[17] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," *CoRR*, vol. abs/2007.09384, 2020. [Online]. Available: https://arxiv.org/abs/2007.09384

[18] X. Lu, W. Diao, Y. Mao, J. Li, P. Wang, X. Sun, and K. Fu, "Breaking immutable: Information-coupled prototype elaboration for few-shot object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1844–1852.

[19] Z. Wang, B. Yang, H. Yue, and Z. Ma, "Fine-grained prototypes distillation for few-shot object detection," 2024. [Online]. Available: https://arxiv.org/abs/2401.07629

[20] S. Xiang, X. Li, J. Ding, S. Chen, and Z. Hua, "Unidirectional local-attention autoencoder network for spectral variability unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[21] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[22] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[23] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3090–3106, 2023.

[24] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 180–10 189.

[25] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7359–7368.

[26] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Mmdetection: Open mmlab detection toolbox and benchmark," 2019. [Online]. Available: https://arxiv.org/abs/1906.07155

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[29] G. Han and S.-N. Lim, "Few-shot object detection with foundation models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 28 608–28 618.