*2025 International Conference on Intelligent Computing*
*July 26-29, Ningbo, China*
https://www.ic-icc.cn/2025/index.php

# CFA-FSOD: Context-aware Feature Aggregation for Few-Shot Object Detection

Huajie Xu[1,2], Haikun Liao[1,2(✉)], Qiukai Huang[1] and Ganxiao Nong[1]

[1] College of Computer and Electronic Information, Guangxi University, Nanning 530004, China
[2] Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004，China
2213301029@st.gxu.edu.cn

**Abstract.** Few-shot object detection (FSOD) aims to detect novel categories from only a few labeled samples. Most of the meta-learning based FSOD methods tend to rely on static support features which lack adaptability to query contexts and have limited representational power, and they often underutilize class-specific features to refine proposals to promote detection performance. To address these challenges, we propose a novel Context-aware Feature Aggregation for FSOD (CFA-FSOD) that enhances interaction in a support-query bidirectional manner. Concretely, in this method, a Query-guided Support Enhancement (QSE) module is proposed to adaptively integrate features from query image regions (typically proposals) into support features to enhance their flexibility; meanwhile, a Cross-attention Feature Modulation (CFM) module is proposed to leverage the enhanced support features to refine query proposals for fine-grained alignment. Experimental results on both Pascal VOC and MS COCO demonstrate that CFA-FSOD achieves outstanding performance in most evaluation settings, benefiting from its bidirectional interaction mechanism that improves the efficiency of sample utilization and the transfer of category-specific features.

**Keywords:** Few-shot Object Detection, Meta Learning, Feature Aggregation

## 1 Introduction

Deep convolutional neural networks (CNNs) have achieved remarkable success in object detection, but their performance heavily depends on large-scale annotated data. In real-world domains such as remote sensing, medical imaging, and industrial inspection, privacy concerns, security restrictions, and high annotation costs often result in insufficient annotated data, which may lead to overfitting and poor generalization in object detectors. To address this challenge, few-shot object detection (FSOD) [1][2][3] has emerged as a promising solution. FSOD aims to pre-train models on extensively annotated base classes and subsequently transfer the learned knowledge to novel classes with only a limited number of labeled samples, enabling accurate detection of objects from novel categories.

Meta-learning has been widely adopted in FSOD to alleviate overfitting caused by limited training data. It follows a task-oriented training paradigm to learn class-agnostic detectors that generalize across tasks. Each task consists of a support set (a few labeled instances) and a query set (unlabeled samples to be detected) in meta-learning, where the detector learns to recognize objects in the query image by exploiting interactions between support and query features. To enhance such interactions, various methods have been proposed. For instance, Multi-Relation Detector [4] performs multi-perspective feature comparison; Meta Faster R-CNN [5] applies spatial alignment to refine class features; and UNP [3] dynamically optimizes gradients using cosine similarity. While effective, these methods typically construct static support features by averaging support samples, resulting in coarse representations that not only lack query-aware information and adaptability to diverse query samples, but also suffer from distributional bias due to the limited and potentially unrepresentative support samples. In addition, these methods often fail to fully exploit class-specific features to refine proposal features, which may limit their overall discriminative capability.

To address the limitations of static support features and insufficient proposal refinement in few-shot object detection, we propose Context-aware Feature Aggregation for FSOD (CFA-FSOD), a novel framework built upon Meta Faster R-CNN that enhances support-query interaction bidirectionally. Specifically, a novel Query-guided Support Enhancement (QSE) module is proposed, which incorporates class features from query proposals into support features by evaluating global and local consistency, yielding more adaptive and representative ones. Furthermore, based on the refined support features, a Cross-attention Feature Modulation (CFM) module is proposed to modulate proposal features in a similarity-aware manner, enabling fine-grained alignment between support features and region proposals. With these modules, CFA-FSOD enhances support features and refines query proposals, effectively improving detection performance.

Our contributions are threefold:

1.We propose CFA-FSOD, a novel FSOD framework that enables bidirectional support-query interaction, addressing limitations of prior methods in static support feature and coarse proposal refinement.

2.Unlike prior FSOD methods, the proposed QSE module evaluates consistency between query proposals and support features at global and local levels to select high-quality proposals, filter out noise, and aggregate their features into support features, thereby improving their adaptability to diverse query samples.

3.We propose the CFM module, which effectively highlights the features in proposals that are highly correlated with class-specific information and fuses them into the corresponding proposals for modulation, thereby enhancing the model's ability to distinguish objects.

## 2 Related work

**Object Detection.** Conventional object detection methods are generally classified into single-stage and two-stage frameworks. Single-stage models, such as the YOLO [6],

utilize a backbone network to extract image features and directly predict object categories and bounding box coordinates. In comparison, two-stage detectors such as the widely used Faster R-CNN [7], a foundation for many FSOD approaches, first generate class-agnostic proposals and then classify and refine them. Despite these advances, most existing detectors encounter difficulties in few-shot scenarios. This is primarily because they rely on large-scale annotated datasets rich in object instances, which are expensive and labor-intensive to obtain in practice.

**Few-Shot Object Detection.** Few-shot object detection (FSOD) aims to detect novel-class objects with only a few annotated samples. Early transfer learning methods often suffered from overfitting, while meta-learning approaches improve generalization by leveraging prior knowledge from base classes. A majority of meta-learning-based FSOD methods focus on modeling interactions between query proposals and class-specific support features. Meta R-CNN [8] aligns proposals with support prototypes through a dedicated support branch. FSDetView [9] fuses support and query features for task-specific prediction. DRL [10] constructs a dynamic graph among proposals to capture their dependencies. Meta Faster R-CNN [5] performs spatial alignment between support and query features, and UNP [3] reweights sample gradients based on proposal-prototype affinity. QA-FewDet [11] propagates information across proposals and support via a heterogeneous graph to enhance feature interaction, while CoAE [12] extracts common semantics across images and maps them into a shared embedding space for better generalization. Inspired by these methods, we propose CFA-FSOD, a framework featuring a bidirectional interaction mechanism between support features and query proposals for more effective class-specific feature interaction and fusion.

# 3 Method

## 3.1 Problem Statement

Following prior work [2][3] in FSOD, we split the dataset into a base dataset $D_b$ with abundant samples and annotations from base classes $C_b$ and a novel dataset $D_n$ with $K$ samples per novel class $C_n$, i.e., $C_b \cap C_n = \emptyset$. The goal of FSOD is to train a detector on $D_b$ that can quickly adapt to $D_n$ in the $K$-shot setting while detecting both base and novel classes. The training process consists of two stages: base training, where the detector is trained on $D_b$, and fine-tuning, where it is finetuned on a balanced dataset $D_{bal}$ formed by combining a subset of $D_b$ with $D_n$ to ensure equal object counts per category. Following meta-learning conventions, both stages use a support set $D_s = \{x^s, y^s\}$ following the $N$-way $K$-shot setting and a query set $D_q = \{x^q, y^q\}$, where in base training both sets come from $D_b$ while in fine-tuning they come from $D_{bal}$. The support set can be formally represented as $D_s = \{x^s_{n,k}, y^s_{n,k}\}$, where $n = 1,2,\dots,N$ indexes the class, $k = 1,2,\dots,K$ indexes the instance. Based on this, query set images are processed using the support set as reference.

Our goal is to enrich static support features with query-aware features derived from high-quality proposals, and leverage them to enhance proposal features for learning more discriminative object representations. As illustrated in Fig. 1, the proposed CFA-

FSOD method involves the following steps: given a query image and *K* support instances per class (*K*-shot), shared backbone networks extract query and support feature maps separately. The *K* support feature maps of each class are averaged to obtain the average support feature map. Meanwhile, the query feature map is passed through a Region Proposal Network (RPN) to generate region proposals, which represent localized object hypotheses and encode the query image's local visual patterns. In the QSE module, the average support feature map and proposal features are compared using a distance function to compute feature consistency scores (FCS), which are then used to weight and aggregate proposal features for modulating the support features. The CFM module then enhances the key features within proposals by leveraging semantic correlations and similarity-conditioned modulation between the modulated support features and the proposals. Finally, the support features and proposal features output separately by the two modules are aggregated and fed into a binary classifier and a regressor to complete the detection.

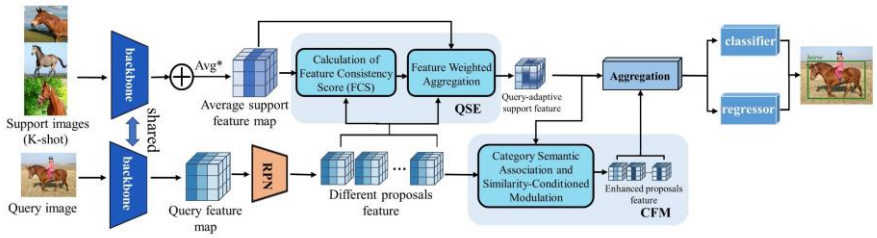


**Fig. 1.** The overall framework of CFA-FSOD.

### 3.2 Query-guided Support Enhancement (QSE)

Meta-learning-based FSOD methods typically average support sample features to form static support features. However, due to the limited number of support samples, such features often exhibit distributional bias and adapt poorly to diverse query samples. Certain methods (e.g., QA-FewDet, CoAE) attempt to address this by propagating proposal features into the support features. Nevertheless, their indiscriminate feature fusion can introduce noise and irrelevant features, ultimately impairing detection accuracy. To address these challenges, as shown in Fig. 2, we design a Query-guided Support Enhancement (QSE) module, which includes two processing steps: Feature Consistency Score, a joint evaluation of global semantics and local context that selects informative proposals, and Feature-Weighted Aggregation, which integrates the features of the selected proposals to refine the support features. This design improves the representativeness of the support features while suppressing the interference of irrelevant features, enabling a shift from static priors to dynamic adaptation.

**Feature Consistency Score.** The classification scores produced by the box classifier are typically used as an implicit measure of proposal localization quality in R-CNN-based detectors. However, since classification scores are primarily influenced by high-confidence regions within the proposal, they often fail to accurately reflect its overall

localization quality. As shown in Fig. 3, some proposals (e.g., the green box) may receive classification scores similar to those of well-localized boxes, even though they are clearly misaligned with the ground truth. While such inconsistencies can be mitigated with large-scale training data, the scarcity of samples in novel categories limits the model's generalization ability, often resulting in low-quality proposals during few-shot detection.

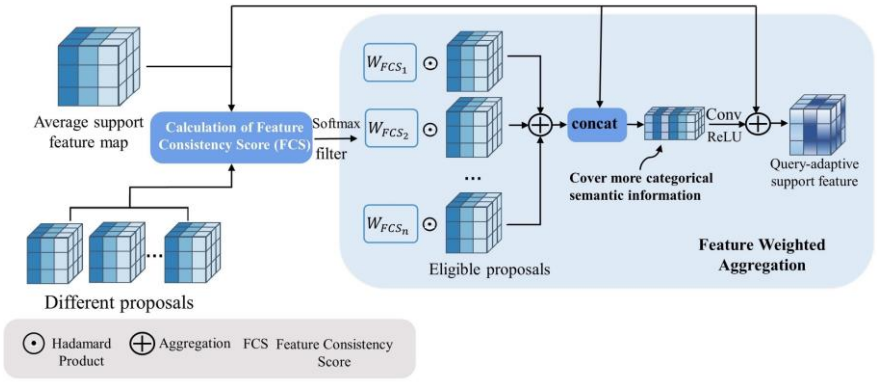


**Fig. 2.** The flowchart of Query-guided Support Enhancement (QSE) Module.

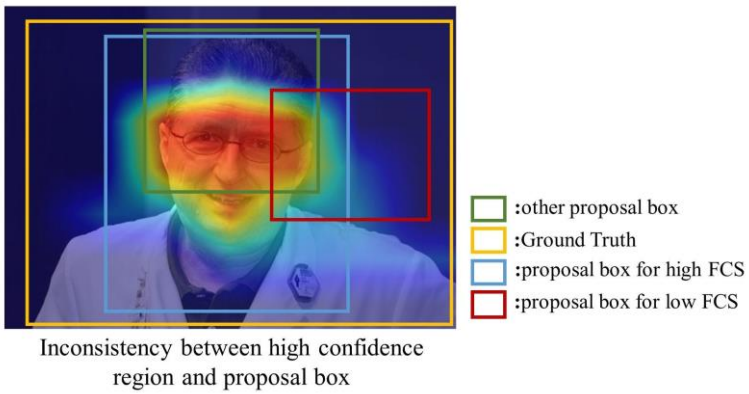


**Fig. 3.** Motivation for evaluating proposal quality using feature consistency scores (FCS). The features of high-FCS proposals are more informative for target representation, as they contain richer local details.

To mitigate the impact of low-quality proposals, we introduce a Feature Consistency Score (FCS) based on a global-local cooperative evaluation mechanism. By measuring semantic and spatial consistency, FCS filters proposals to retain those with accurate classification and localization (e.g., the blue box in Fig. 3). Features of these selected proposals are then dynamically aggregated to enhance the adaptability of the support features.

Specifically, given a support set $S_c$ of target class $c$ and a proposal feature map $r \in \mathrm{R}^{C\times H\times W}$, we first obtain the class-wise average support feature map $\bar{x}_c^s \in \mathrm{R}^{C\times H\times W}$ by computing the mean of all support feature maps in $S_c$, which is then used to compute the FCS for each proposal as follows:

$$FCS(\bar{x}_c^s, r) = \sigma\left[(1-\alpha)D\left(v_{s,c}, v_r\right) + \alpha D\left(f_{s,c}, f_r\right)\right] \tag{1}$$

$$D(x, y) = \frac{x^T y}{||x|| \cdot ||y||} \tag{2}$$

$$\sigma(x) = \frac{1}{1 + e^{-\lambda x}} \tag{3}$$

Where $\alpha$ is a balancing coefficient, and $D(\cdot)$ and $\sigma(\cdot)$ denote the cosine similarity and the sigmoid function with scaling factor $\lambda$, respectively. Since $\sigma(\cdot)$ is strictly monotonic, $\lambda$ (empirically set to 3) only controls the sharpness of the score distribution without affecting ranking. $v_{s,c}$, $v_r \in \mathrm{R}^C$ denote the global semantic feature vectors of the support and proposal, respectively, obtained by applying global average pooling (GAP) to $\bar{x}_c^s$ and $r$. These vectors encode semantic information useful for classification. $f_{s,c}$, $f_r \in \mathrm{R}^{CHW}$ denote the flattened local feature vectors of the support and proposal, respectively, obtained by reshaping $\bar{x}_c^s$ and $r$. Although flattening discards explicit spatial coordinates, variations in local responses still implicitly reflect spatial consistency. This dual-level comparison jointly evaluates semantic similarity and spatial consistency for robust proposal scoring.

**Feature-Weighted Aggregation.** After computing the FCS for all proposals, a preset threshold $\tau$ is applied to filter proposals, retaining only those proposals satisfy $FCS(\bar{x}_c^s, r) \geq \tau$. This deterministic filtering mechanism ensures that subsequent feature fusion focuses on proposals with strong semantic relevance to the support features. Then, normalized weights are assigned to the selected proposals according to their FCS values, as follows:

$$w_{FCS_i} = \frac{FCS(\bar{x}_c^s, r_i)}{\sum_{j=1}^{n} FCS\left(\bar{x}_c^s, r_j\right)} \tag{4}$$

Where $n$ denotes the number of proposals that satisfy the filtering criterion; $r_i$ and $w_{FCS_i}$ represent the $i$-th filtered proposal and its corresponding normalized weight, respectively. Based on these normalized weights, the query-aware feature $r^q \in \mathrm{R}^{C\times H\times W}$ is constructed via weighted aggregation, as follows:

$$r^q = \sum_i^n w_{FCS_i} \cdot r_i \tag{5}$$

Finally, the original support feature $\bar{x}_c^s$ and the query-aware feature $r^q$ are concatenated and passed through a 1×1 convolution followed by *ReLU* activation. The resulting feature is then element-wise added to $\bar{x}_c^s$ to preserve the original feature and produce the query-adaptive support feature $\acute{x}_c^s \in \mathrm{R}^{C\times H\times W}$, as follows:

$$\acute{x}_c^s = \bar{x}_c^s + ReLU\left(Conv(Concat[\bar{x}_c^s; r^q])\right) \tag{6}$$

### 3.3 Cross-attention Feature Modulation (CFM)

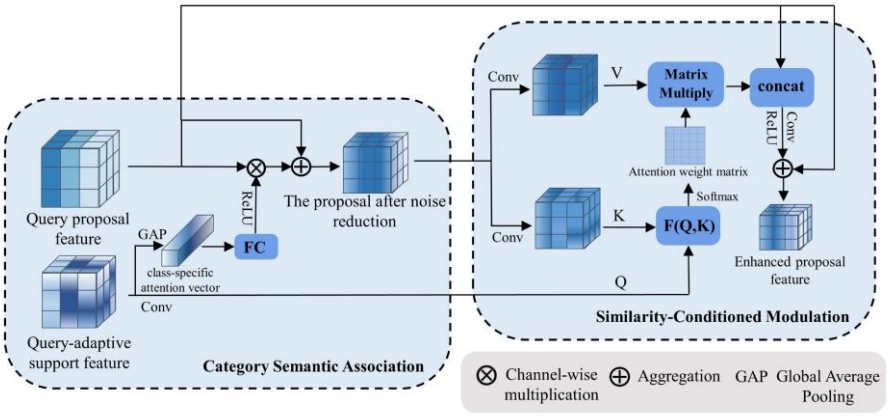


**Fig. 4.** The flowchart of Cross-attention Feature Modulation (CFM) Module.

Effective support-query interactions play a crucial role in transferring class-specific features in meta-learning based FSOD. When proposal features lack fine-grained alignment with support features, these interactions weaken, reducing the model's ability to distinguish objects from background and suppress irrelevant regions. This issue impairs the accurate localization and classification of objects, especially when they are occluded or deformed. To address this, a Cross-attention Feature Modulation (CFM) module is proposed, as illustrated in Fig. 4, which performs non-local interactions between the support features enhanced by the QSE module and the query proposal features to capture pixel-level similarities. These similarities are then employed to conditionally modulate the proposal features, enabling more effective feature refinement. The CFM module consists of two components: Category Semantic Association (CSA) and Similarity-Conditional Modulation (SCM), described as follows.

**Category Semantic Association.** When the feature map has a high channel dimensionality, noisy or irrelevant features may hinder the non-local interaction mechanism from effectively focusing on key features. To mitigate this issue, we introduce a channel attention mechanism that uses the query-adaptive support feature $\acute{x}_c^s$ as a semantic prior to suppress weakly relevant feature responses. Specifically, a class-specific attention vector is first obtained by applying global average pooling (GAP) to $\acute{x}_c^s$. This vector is then transformed through a fully connected (*FC*) layer followed by a *ReLU* activation to introduce non-linearity, thereby yielding channel-wise attention weights that are then applied to the proposal feature $r$ to suppress channels that are less relevant to the target class *c*. To preserve the original feature, the weighted features are finally added back to $r$, as defined in Eq. (7), where $\otimes$ denotes channel-wise multiplication.

$$r = r + r \otimes ReLU\big(FC(GAP(\acute{x}_c^s))\big) \tag{7}$$

**Similarity-Conditioned Modulation.** For conveying class-specific features from the query-adaptive support feature $\acute{x}_c^s$ to the proposal $r$, we compute a similarity matrix based on a query-key compatibility function [13] to model their global dependencies.

Specifically, $\acute{x}_c^s$ is first projected by a 1×1 convolution to produce the query $Q \in \mathrm{R}^{C' \times H \times W}$, while $r$ is transformed into the key $K \in \mathrm{R}^{C' \times H \times W}$ and value $V \in \mathrm{R}^{C' \times H \times W}$ via two separate 1×1 convolutions. These projections reduce channel dimensionality and enhance inter-channel interactions. Then, $Q$ and $K$ are reshaped into matrices of size $C' \times HW$, and the dot-product similarity matrix $Q^T K \in \mathrm{R}^{HW \times HW}$ is computed to quantify the pairwise spatial similarity between the transformed support and proposal features. Finally, a row-wise SoftMax is applied to $Q^T K$ to obtain the normalized attention weight matrix $F(Q,K) \in \mathrm{R}^{HW \times HW}$, as follows:

$$F(Q,K) = softmax(Q^T K) \tag{8}$$

Where $F(Q,K)_{i,j}$ denotes the attention weight between the $i$-th spatial location in $Q$ and the $j$-th spatial location in $K$. For matrix multiplication, $V$ is reshaped into matrices of size $HW \times C'$, and is then multiplied by $F(Q,K)$ to aggregate features across spatial locations based on the learned attention weights, highlighting features highly correlated with class-specific information, as follows:

$$\hat{r} = F(Q,K) \cdot V \tag{9}$$

Where $\hat{r} \in \mathrm{R}^{HW \times C'}$ denotes the attention-weighted modulated feature, which is then reshaped back into a feature map of size $C' \times H \times W$ for subsequent feature fusion.

To integrate the modulated feature $\hat{r}$ with the original proposal feature $r$, we concatenate $r$ and $\hat{r}$ along the channel dimension and apply a 1×1 convolution followed by a *ReLU* activation to adaptively fuse the refined feature. Finally, to preserve the original information, the fused output is added back to $r$, yielding the final proposal representation $r' \in \mathrm{R}^{C \times H \times W}$, as follows:

$$r' = r + ReLU(Conv(Concat[r;\hat{r}])) \tag{10}$$

# 4 Experiments

## 4.1 Dataset and Experimental Details

**Benchmark Datasets.** Following prior work [3][5], we evaluate our method on the PASCAL VOC [14] and MS COCO [15] datasets. For PASCAL VOC, the model is trained on the combined trainval sets of VOC 2007 and 2012 [16], and tested on the VOC 2007 test set. Three few-shot splits are defined by selecting 5 novel and 15 base classes, with $K$ = {1, 2, 3, 5, 10} annotated instances per novel class used for fine-tuning. For MS COCO, 20 categories overlapping with PASCAL VOC are treated as novel classes and the remaining 60 as base classes. Evaluation is performed with $K$ = {10, 30} instances per novel class randomly sampled for fine-tuning.

**Experimental Setup.** We adopt an episodic training scheme, where each query image is paired with a 2-way 30-shot support set during base training. Support images are constructed by cropping ground-truth objects and resized to 320×320 pixels. Training consists of two stages: base training and fine-tuning. Base training includes three phases with a batch size of 8. For PASCAL VOC, base training is divided into three phases with a batch size of 8. The learning rate is set to 0.002 for the first 20k iterations (phase 1), followed by two phases of 10k iterations each (phases 2 and 3) with a reduced learning rate of 0.001. For MS COCO, the same settings are used, but the iteration count is

doubled to accommodate the larger dataset. In the fine-tuning stage, following the TFA [17] protocol, the backbone is frozen while the detection head is updated using a few novel-class samples. The batch size, learning rate, and iteration count are uniformly set to 8, 0.001, and 3,000 for both datasets.

**Evaluation Metrics.** Consistent with most FSOD studies [5][6][10], we report nAP50 (average precision of novel classes at IoU = 0.5) on the PASCAL VOC 2007 test set. For MS COCO, we adopt the standard evaluation metrics nAP, nAP50, and nAP75, which represent the average precision of novel classes at IoU thresholds of [0.5:0.95], 0.5, and 0.75, respectively. Notably, nAP75 imposes a stricter matching criterion than nAP50, requiring a higher overlap between predicted and ground-truth (GT) boxes.

### 4.2 Performance Comparisons

We evaluated the proposed method against the mainstream state-of-the-art methods, including meta-learning-based approaches (e.g., Meta R-CNN [8], FSDetView [9], QA-FewDet [11], DRL [10], Meta Faster R-CNN [5], UNP [3]) and transfer-learning-based methods (e.g., TFA [17], FSCE [18], DiGeo [19], FS$^3$C [1], FSRC [20], FSNA [21], EME [22]), following standard evaluation protocols. For fairness, all methods were implemented under the Faster R-CNN framework with a ResNet-101 [23] backbone, consistent with most FSOD studies. In the results table, bold indicates the best result, underline marks the second-best, '†' denotes a reproduced baseline under identical settings, and '-' indicates missing values in prior literature.

**Table 1.** The comparative experimental results on PASCAL VOC dataset (%)

| Method/Shots | Novel Split 1 | | | | | Novel Split 2 | | | | | Novel Split 2 | | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | |
| MetaRCNN | 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 | 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 |
| FSDetView | 24.3 | 35.3 | 42.2 | 49.1 | 57.4 | 21.6 | 24.6 | 31.9 | 37.0 | 45.7 | 21.2 | 30.0 | 37.2 | 43.8 | 49.6 | 36.7 |
| FSCE | 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 | 46.6 |
| QA-FewDet | 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 | 48.0 |
| DRL | 28.0 | 40.5 | 49.4 | 49.9 | 59.4 | 22.9 | 33.4 | 36.4 | 36.1 | 52.7 | 28.0 | 32.0 | 40.4 | 46.7 | 53.5 | 40.6 |
| DiGeo | 37.9 | 39.4 | 48.5 | 58.6 | 61.5 | 26.6 | 28.9 | 41.9 | 42.1 | 49.1 | 30.4 | 40.1 | 46.9 | 52.7 | 54.7 | 44.0 |
| FS$^3$C | 44.1 | 46.5 | 51.6 | 57.8 | 61.5 | 27.6 | 29.5 | 40.9 | 40.0 | 46.8 | 40.4 | 44.5 | 46.0 | 52.9 | 55.6 | 45.7 |
| FSRC | 45.5 | 43.4 | 51.1 | 61.4 | 64.0 | 28.4 | 31.1 | 45.0 | 46.1 | 51.6 | 38.8 | 45.1 | 48.4 | 55.5 | 59.0 | 47.6 |
| FSNA | 43.8 | 47.7 | 50.8 | 57.4 | 60.3 | 23.9 | 32.3 | 37.9 | 40.2 | 41.8 | 34.0 | 40.7 | 45.5 | 52.3 | 54.0 | 44.2 |
| UNP | 43.7 | 58.3 | 59.8 | 63.7 | 64.2 | 28.1 | 42.8 | **47.7** | 49.5 | 50.3 | 38.4 | **49.3** | 53.8 | 57.7 | 58.7 | 51.1 |
| EME | 41.8 | 47.1 | 49.6 | 58.6 | 62.6 | **29.6** | 31.2 | 40.6 | 44.4 | 47.7 | 36.6 | 40.9 | 46.1 | 49.7 | 52.7 | 45.3 |
| Meta FRCNN† | 43.1 | 54.9 | 59.0 | 63.2 | 65.8 | 27.3 | 35.8 | 44.3 | 48.1 | 52.1 | 40.1 | 47.3 | 53.4 | 59.1 | 58.6 | 50.1 |
| CFA-FSOD(Ours) | **48.0** | **58.7** | **61.9** | **65.8** | **67.4** | 29.2 | **43.2** | 47.0 | **51.8** | **52.8** | **41.0** | **49.3** | **54.2** | **59.4** | **59.8** | **52.6** |

**PASCAL VOC.** Table 1 summarizes the detection results of the proposed method CFA-FSOD compared to the mainstream state-of-the-art approaches on the PASCAL

VOC dataset. CFA-FSOD consistently outperforms the Meta Faster R-CNN baseline, achieving average improvements of 2.6%, 4.4%, 2.1%, 2.2%, and 1.2% across the 1-, 2-, 3-, 5-, and 10-shot settings. Notably, the performance improvements tend to be more pronounced in scenarios with fewer shots. This can be attributed to the Query-guided Support Enhancement (QSE) module, which effectively alleviates the distributional bias of support features caused by limited support samples by aggregating high-quality proposal features from the query image. Compared with other SOTA methods, CFA-FSOD achieves the highest nAP50 in 13 out of 15 trials (3 splits × 5 shot settings) and ranks second in the other 2 trials, demonstrating its superiority in novel class detection.

**MS COCO.** We further evaluated the proposed method on the more challenging MS COCO dataset, with results summarized in Table 2. Compared to the baseline Meta Faster R-CNN, our method achieves AP improvements of 1.4% and 1.2% in the 10-shot and 30-shot settings, respectively. Furthermore, our method consistently outperforms other SOTA approaches on most evaluation metrics, confirming that CFA-FSOD remains effective on challenging and complex datasets.

**Table 2.** The comparison experimental results on MS COCO dataset (%)

| Method/Shots | 10-shot | | | 30-shot | | |
|---|---|---|---|---|---|---|
| | nAP | nAP50 | nAP75 | nAP | nAP50 | nAP75 |
| Meta R-CNN | 8.7 | 19.1 | 6.6 | 12.4 | 25.3 | 10.8 |
| TFA w/cos | 10.0 | 17.1 | 8.8 | 13.7 | 22.0 | 12.0 |
| FSDetView | <u>12.5</u> | **27.3** | 9.8 | 14.7 | 30.6 | 12.2 |
| DRL | 10.9 | 25.2 | 7.0 | 15.0 | 31.7 | 11.8 |
| FSCE | 11.9 | - | 10.5 | 16.4 | - | <u>16.2</u> |
| QA-FewDet | 11.6 | 23.9 | 9.8 | <u>16.5</u> | <u>31.9</u> | 15.5 |
| DiGeo | 10.3 | 18.7 | 9.9 | 14.2 | 26.2 | 14.8 |
| FS$^3$C | 11.0 | 23.6 | 10.0 | 15.1 | 28.9 | 14.9 |
| FSRC | 12.0 | - | 10.7 | 16.4 | - | 15.7 |
| FSNA | 11.9 | 25.4 | 10.3 | 16.1 | 31.1 | 15.1 |
| UNP | 12.3 | 23.1 | <u>11.5</u> | 15.3 | 28.5 | 14.8 |
| EME | 10.6 | 19.8 | 10.2 | 15.1 | 27.3 | 15.7 |
| Meta FR-CNN† | 12.3 | 25.1 | 10.6 | 16.0 | 31.2 | 14.8 |
| CFA-FSOD(Ours) | **13.7** | <u>26.7</u> | **12.0** | **17.2** | **32.5** | **16.3** |

### 4.3 Ablation Studies

To verify the effectiveness of the proposed modules, we conducted a series of ablation studies. First, we evaluated the individual contributions of the Query-guided Support Enhancement (QSE) and Cross-attention Feature Modulation (CFM) modules to the overall CFA-FSOD framework. Next, we investigated the effect of the balancing coefficient $\alpha$ and threshold $\tau$ in the QSE module. Finally, we analyzed the impact of the projection channel dimension used in computing $Q$, $K$, and $V$ within the CFM module.

**Ablation of different modules.** Table 3 demonstrates the effectiveness of CFA-FSOD's modules. The QSE module alone brings a 2.3% average improvement, demonstrating that incorporating class features from query proposals into static support features improves their adaptability. For the CFM module, introducing the Similarity-Conditioned Modulation (SCM) component yields a 0.8% average gain by capturing class-relevant features through non-local interactions. Adding the Category Semantic Association (CSA) component on top of SCM further boosts the gain to 1.5%, as semantic priors help suppress noise and improve feature quality. Combining QSE with CFM achieves a 3.2% average gain, indicating that the synergy between the two modules facilitates more effective fine-grained alignment between support features and proposal features.

**Analysis of the balancing coefficient $\alpha$.** We analyze the impact of the balancing coefficient $\alpha$ in the Feature Consistency Score (FCS), which guides the QSE module in selecting high-quality proposals. As shown in Table 4, intermediate values of $\alpha$ consistently yield better results than extreme settings ($\alpha$ = 0.0 or 1.0), confirming the complementary nature of global semantics and local contextual patterns. The best performance occurs at $\alpha$ = 1/4, where spatial consistency is moderately emphasized without compromising the global semantic features' ability to convey category information. In contrast, larger $\alpha$ values emphasize local contextual information through flattened local feature vectors, which may degrade the accuracy of cosine similarity due to their high dimensionality and sensitivity to spatial noise.

**Table 3.** Ablation experiment results of the QSE and CFM modules (%)

| No | QSE | CFM | | Shots | | | | | *Avg* | $\Delta Avg$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CSA | SCM | 1 | 2 | 3 | 5 | 10 | | |
| 1 | × | × | × | 43.1 | 54.9 | 59.0 | 63.2 | 65.8 | 57.2 | - |
| 2 | ✓ | × | × | 46.8 | 57.3 | 61.0 | 65.2 | 67.2 | 59.5 | +2.3 |
| 3 | × | × | ✓ | 44.6 | 55.4 | 59.6 | 64.0 | 66.5 | 58.0 | +0.8 |
| 4 | × | ✓ | ✓ | 45.9 | 56.2 | 60.4 | 64.4 | 66.8 | 58.7 | +1.5 |
| 5 | ✓ | ✓ | ✓ | **48.0** | **58.7** | **61.9** | **65.8** | **67.4** | **60.4** | **+3.2** |

**Table 4.** Analysis of the balancing coefficient $\alpha$ in the QSE module (%)

| No | $\alpha$ | Shots | |
|---|---|---|---|
| | | 3 | 10 |
| 1 | 0 | 59.9 | 66.3 |
| 2 | 1/4 | **61.0** | **67.2** |
| 3 | 1/2 | 60.6 | 67.0 |
| 4 | 3/4 | 60.0 | 66.5 |
| 5 | 1 | 59.7 | 66.4 |

**Analysis of the threshold $\tau$.** We analyze the impact of the threshold $\tau$ for selecting high-quality proposals. As shown in Table 5, the best performance is achieved at $\tau$ =0.8, which effectively balances filtering low-quality proposals and preserving informative regions. A lower threshold ($\tau$ =0.7) allows noisy proposals, while a higher threshold

($\tau$ =0.9) may discard complementary proposals that help cover the full extent of the target object, leading to incomplete query-aware features.

**Table 5.** Analysis of the threshold $\tau$ in the QSE module. (%)

| No | $\tau$ | Shots | |
|---|---|---|---|
| | | 3 | 10 |
| 1 | 0.7 | 59.7 | 66.3 |
| 2 | 0.8 | **61.0** | **67.2** |
| 3 | 0.9 | 60.5 | 66.8 |

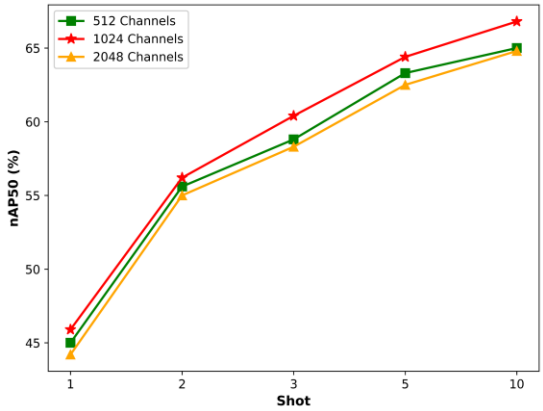


**Fig. 5.** Analysis of the projection dimension of *Q*, *K*, and *V* in the CFM module.

**Analysis of the projection dimension in CFM module.** We analyze the impact of the intermediate channel dimension $C'$ used to project the query (*Q*), key (*K*), and value (*V*) in our CFM module. This dimension controls the representation capacity of the attention mechanism. As shown in Fig. 5, using 1024 channels achieves the best performance. Reducing it to 512 limits the feature representation capacity of the feature maps, while increasing it to 2048 may overfit noise from limited training samples due to the high-dimensional feature projections. We use 1024 as the default setting for a good trade-off between accuracy and generalization.

### 4.4 Visualization

In Fig. 6, we visualize the heatmaps generated by the baseline, the CFM module alone, and the complete CFA-FSOD framework (i.e., CFM + QSE). The results show that CFM helps the model focus on target-relevant regions. With the integration of QSE, CFA-FSOD further refines the attention by incorporating query-adaptive support features, enabling the model to concentrate more precisely on category-relevant regions. These results suggest that the bidirectional interaction between support and query features contributes to improving the model's ability to extract key features during the target recognition process. We further present the detection results of our method and

the baseline in Fig. 7. The proposed CFA-FSOD demonstrates superior performance, effectively reducing false positives and false negatives in challenging scenarios such as appearance variations and occlusions. This validates the effectiveness of the overall design, where the CFM module employs cross-attention to align the query proposal features with the support features enhanced by the QSE module, thereby improving the model's perception of foreground objects.

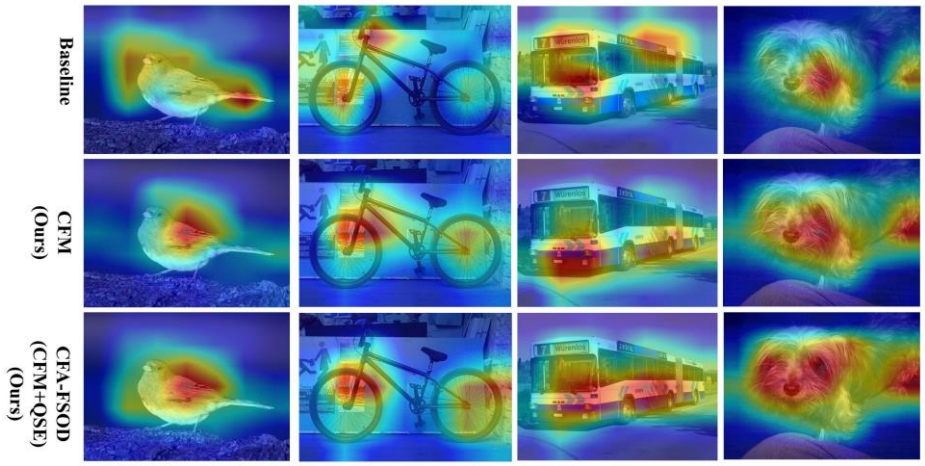


**Fig. 6.** Qualitative Heatmap Results on the VOC Dataset. The first three rows correspond to the heatmaps generated by the baseline, CFM module, and the complete CFA-FSOD framework, respectively.


**Fig. 7.** Detection results of the baseline and our method CFA-FSOD.

## 5 Conclusion

In this paper, by addressing the limitations of static support features and insufficient proposal refinement in most of the meta-learning based FSOD methods, we propose Context-aware Feature Aggregation for FSOD (CFA-FSOD), a novel framework for few-shot object detection by enhancing interaction in a support-query bidirectional manner. Within this framework, a Query-guided Support Enhancement (QSE) module

is proposed to dynamically adapt support features by evaluating proposal-support consistency, while a Cross-attention Feature Modulation (CFM) module is proposed to refine proposals through similarity-guided attention. Experiments on PASCAL VOC and MS COCO demonstrate that CFA-FSOD consistently outperforms most existing state-of-the-art methods, confirming the effectiveness of bidirectional support-query interaction in improving few-shot detection. In our future work, the support features are constructed adaptively by considering the relative importance of support samples to improve generalization to diverse query samples.

**Acknowledgments.** The authors greatly acknowledge the financial support from the Natural Science Foundation of Guangxi Zhuang Autonomous Region (Grant No. 2024JJA170106), the Key Research and Development Program of Guangxi (Grant No. AD25069071), and the National Natural Science Foundation of China (Grant No. 52169021).

**Disclosure of Interests.** The authors declare no conflict of interest.

## References

1. Qi, D., Hu, J., Shen, J.: Few-Shot Object Detection with Self-Supervising and Cooperative Classifier. IEEE Trans. Neural Networks Learn. Syst. **35**(4), 5435-5446 (2024)
2. Chen, H., Wang, Q., Xie, K., et al.: SD-FSOD: Self-Distillation Paradigm via Distribution Calibration for Few-Shot Object Detection. IEEE Trans. Circuits Syst. Video Technol. **34**(7), 5963-5976 (2024)
3. Yan, B., Lang, C., Cheng, G., et al.: Understanding Negative Proposals in Generic Few-Shot Object Detection. IEEE Trans. Circuits Syst. Video Technol. **34(**7), 5818-5829 (2024)
4. Fan, Q., Zhuo, W., Tang, C., et al.: Few-shot object detection with attention-RPN and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4013-4022 (2020)
5. Han, G., Huang, S., Ma, J., et al.: Meta Faster R-CNN: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 780-789 (2022)
6. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, pp. 21-37 (2016)
7. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137-1149 (2016)
8. Yan, X., Chen, Z., Xu, A., et al.: Meta R-CNN: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9577-9586 (2019)
9. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3090-3106 (2022)
10. Liu, W., Cai, X., Wang, C., et al.: Dynamic relevance learning for few-shot object detection. Signal Image Video Process **19**(4), 297 (2025)
11. Han, G., He, Y., Huang, S., et al.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3263-3272 (2021)
12. Hsieh, T., Lo, Y., Chen, H., et al.: One-shot object detection with co-attention and co-excitation. In: Advances in Neural Information Processing Systems, vol. 32, pp. 560-568 (2019)

13. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 30 (2017)
14. Everingham, M., Van, G., Williams, C., et al.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303-338 (2010)
15. Lin, T., Maire, M., Belongie, S., et al.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 740-755 (2014)
16. Everingham, M., Eslami, S., Van, G., et al.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. **111**, 98-136 (2015)
17. Wang, X., Huang, T., Darrell, T., et al.: Frustratingly Simple Few-Shot Object Detection. In: 37th International Conference on Machine Learning: ICML, pp. 9919-9928 (2021)
18. Sun, B., Li, B., Cai, S., et al.: FSCE: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7352-7362 (2021)
19. Ma, J., Niu, Y., Xu, J., et al.: DiGeo: Discriminative geometry-aware learning for generalized few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3208-3218 (2023)
20. Shangguan, Z., Huai, L., Liu, T., et al.: Few-shot object detection with refined contrastive learning. In: IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 991-996 (2023)
21. Zhu, J., Wang, Q., Dong, X., et al.: FSNA: Few-Shot Object Detection via Neighborhood Information Adaption and All Attention. IEEE Trans. Circuits Syst. Video Technol. **34**(8), 7121-7134 (2024)
22. Liu, C., Li, B., Shi, M., et al.: Explicit margin equilibrium for few-shot object detection. IEEE Trans. Neural Netw. Learn. Syst. **36**(5), 8072-8084 (2025)
23. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016)