Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063109

ARTICLE

Implicit Feature Contrastive Learning for Few-Shot Object Detection

Gang Li[1,#], Zheng Zhou[1,#], Yang Zhang[2,*], Chuanyun Xu[2], Zihan Ruan[1], Pengfei Lv[1], Ru Wang[1], Xinyu Fan[1] and Wei Tan[1]

[1]School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 401331, China
[2]School of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China

*Corresponding Author: Yang Zhang. Email: zhangyang@cqnu.edu.cn

[#]These authors contributed equally to this work

Received: 05 January 2025; Accepted: 29 April 2025; Published: 09 June 2025

ABSTRACT: Although conventional object detection methods achieve high accuracy through extensively annotated datasets, acquiring such large-scale labeled data remains challenging and cost-prohibitive in numerous real-world applications. Few-shot object detection presents a new research idea that aims to localize and classify objects in images using only limited annotated examples. However, the inherent challenge in few-shot object detection lies in the insufficient sample diversity to fully characterize the sample feature distribution, which consequently impacts model performance. Inspired by contrastive learning principles, we propose an Implicit Feature Contrastive Learning (IFCL) module to address this limitation and augment feature diversity for more robust representational learning. This module generates augmented support sample features in a mixed feature space and implicitly contrasts them with query Region of Interest (RoI) features. This approach facilitates more comprehensive learning of both intra-class feature similarity and inter-class feature diversity, thereby enhancing the model's object classification and localization capabilities. Extensive experiments on PASCAL VOC show that our method achieves a respective improvement of 3.2%, 1.8%, and 2.3% on 10-shot of three Novel Sets compared to the baseline model FPD.

KEYWORDS: Few-shot learning; object detection; implicit contrastive learning; feature mixing; feature aggregation

1 Introduction

Object detection constitutes a core computer vision task that combines localization and classification of objects within images with deep learning-based approaches, demonstrating significant advancement in recent years [1–4]. The essence of object detection is to identify and locate the target by discriminating the features of image data. Region of Interest(RoI) can help the model narrow the detection range and focus on the key areas to enhance detection accuracy and efficiency. However, Conventional object detection methodologies require substantial annotated datasets to achieve optimal performance. This requirement poses considerable challenges in real-world applications where limited availability, high costs, or practical difficulties often constrain data acquisition. These limitations have catalyzed the emergence of Few-Shot Object Detection (FSOD).

FSOD is designed to achieve two critical tasks: classifying target objects and precisely localizing them with only a few annotated samples. The inherent scarcity of feature space makes FSOD particularly challenging. The paradigm assumes abundant samples for base classes, whilst novel classes only have a few

examples. Therefore, the main research direction in the field of FSOD focuses on training highly generalizable detection models using limited labeled samples.

There are three primary types of FSOD methods: Transfer learning-based methods [5–7], which leverage pre-trained models from large-scale datasets to augment task-specific learning efficiency. These methods employ parameter freezing [8–10] and gradient decoupling techniques [11,12] during fine-tuning to adapt to few-shot scenarios whilst mitigating overfitting. Data augmentation-based methods [13,14] augment model generalization capabilities. Meta-learning-based methods [8,15–17] follow the 'learning to learn' paradigm by training task-level meta-learners to acquire optimal initial parameters for rapid adaptation to novel tasks.

This paper studies meta-learning-based FSOD, aiming to learn the feature relationship deeply through implicit contrast learning modules and augment the feature diversity by using feature mixing, which has solved the problem of sparse feature space in FSOD.

Our research identifies feature aggregation as an implicit contrastive learning method. As illustrated in Fig. 1, contrastive learning is a feature learning approach that constructs positive and negative sample pairs to learn feature relationships, thereby enhancing model discriminative capability. The feature aggregation approach illustrated in Fig. 2 performs comparisons by selecting both same-class and different-class samples from the support images relative to each query image. This process minimizes distances between same-class samples whilst maximizing distances between different-class samples in feature space, thereby improving model discrimination. This feature aggregation method aligns with the fundamental principles of contrastive learning, allowing us to conceptualize the feature aggregation of positive-negative samples with query RoI as an implicit contrastive learning approach. Furthermore, the amount of training samples constrains the ability of the model to learn the richness and high diversity of feature representations, which leads to the object detection ability not being good enough. To help models learn powerful feature representations, Meta R-CNN [16] employs a straightforward class-specific aggregation method, aggregating the ROI with support sample features from the same class. Variational Feature Aggregation(VFA) [17] designs a class-agnostic aggregation approach that aggregates RoI features with randomly selected support sample features to reduce class bias. Fine-grained prototype distillation(FPD) [18] further develops this concept through its Balanced Class-Agnostic Sampling (B-CAS) strategy, which simultaneously aggregates RoI features with a pair of support samples. This approach prevents the potential pitfalls of random sampling and ensures that the model focuses on critical positive prototypes. However, object detection typically requires a denser feature space to capture the diversity of object appearances, scales, and backgrounds to improve accuracy. When the amount of samples is insufficient, the model cannot learn rich semantic information from a sparse feature space, severely compromising its performance. Therefore, in few-shot scenarios, we argue that the feature space defined by a single positive and negative sample pair is too sparse to learn robust feature representations, resulting in suboptimal performance.

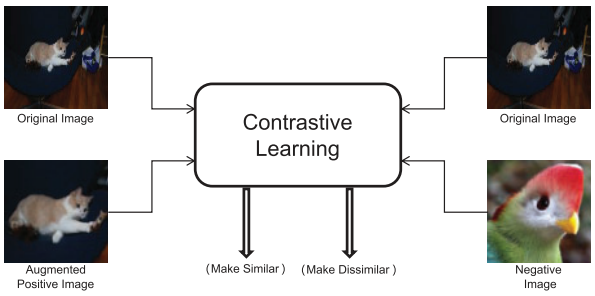

Figure 1: Contrastive Learning paradigm: push similar images closer and push different images away

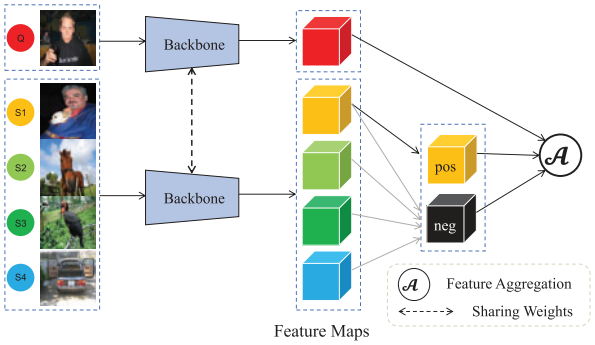

Figure 2: Feature Aggregation method: generate positive and negative samples from the support samples and perform feature aggregation with the query RoI

We propose an Implicit Feature Contrastive Learning (IFCL) module to handle this limitation and help the model learn feature representations. This novel approach incorporates feature mixing techniques to generate augmented features and construct positive-negative sample pairs within an augmented feature space. The subsequent aggregation of these pairs with RoI features strengthens the diversity and robustness of feature representations, effectively improving model performance in few-shot scenarios.

Our contributions are as follows:

1. In this study, inspired by contrastive learning, we analyze the feature aggregation method and find that it can essentially be viewed as an implicit contrastive learning approach. Constructing positive and negative sample pairs and making comparisons helps the model more effectively distinguish novel classes, thereby significantly improving classification and detection performance.

2. We propose a new approach called Implicit Feature Contrastive Learning. Firstly, we employ feature mixing techniques to generate augmented positive and negative sample features. These features are then implicitly contrasted, through feature aggregation, with RoI features in the mixed feature space. The goal is to learn robust feature relationships. Ultimately, this helps the model improve classification and detection accuracy.

3. Through experiments on the widely used PASCAL VOC dataset for object detection, we have proved the efficacy of the proposed method in FSOD tasks. The experimental results demonstrate superior accuracy relative to existing methods. These experiments validate the theoretical analysis and offer experimental evidence for advancing FSOD, showcasing the potential advantages of implicit contrastive learning and feature mixing.

2 Related Work

2.1 Few-Shot Object Detection

FSOD presents more significant challenges compared to few-shot image classification. Current two-stage few-shot object detection methodologies predominantly encompass three categories: data augmentation-based, transfer learning-based, and meta-learning-based methods [19].

2.1.1 Data Augmentation Based Methods

These methods address the limited sample issue by expanding the dataset to augment model generalization. For object detection tasks, two augmentation strategies are employed: those requiring bounding box modifications (e.g., cropping, rotation) and those that are not (e.g., color transformation, noise injection). Wu

et al. [20] augmented data diversity through multi-scale positive samples within feature pyramid networks. Zhang et al. [21] improved performance under extreme data scarcity using virtual samples, whilst Riou et al. [22] implemented sample replication. In addition, semantic embedding [23,24] and augmentation techniques [25–27] have been successfully integrated into a few-shot detection framework.

2.1.2 Transfer Learning Based Methods

Compared to data augmentation-based methods, the transfer learning method does not require additional data collection, has a simple training strategy, and can be used as an effective FSOD method. It does not need to design training tasks but to transfer the base-class-trained detection model to novel classes via fine-tuning. This method does not require a strong correlation between tasks and emphasizes the performance of the new tasks transferred. Combining the advantages of the single-stage detection model SSD [2] and the two-stage detector Faster R-CNN [28], Chen et al. propose a few-shot transfer detector LSTD [29], which can be used to detect unseen objects. Wang et al. proposed TFA [5], which uses a classifier based on cosine similarity to fine-tune the last layer using only new samples to achieve results comparable to other methods. Qiao et al. proposed DeFRCN [7] that uses gradient decoupling technology under the Faster R-CNN framework and refines classification results, improving performance.

2.1.3 Meta-Learning Based Methods

Meta-learning is an encouraging research framework for FSOD. The FSRW [8] proposed by Kang et al. uses a reweighted vector to reweight the YOLOv2 feature map along channel dimensions, which can highlight relevant features. Yan et al. proposed Meta R-CNN [16], which uses the R-CNN framework to construct a twinning network based on double branches. It aggregates queries and supports images to generate RoI features and class prototypes, then fuses them. Han et al. [30] built a meta-classifier through feature alignment and nonlinear matching to replace the conventional softmax-based classifier. It assesses the similarity between the query and support features to produce binary classification outcomes on unseen classes. Zhang et al. [31] designed the Meta-DETR model, which operates at the image level and does not rely on region proposals, thereby avoiding the limitation of inaccurate bounding boxes commonly encountered in FSOD frameworks. The model also processes several support classes simultaneously in a single forward propagation to get inter-class correlations between different classes. However, introducing transformers for constructing the encoder and decoder results in significant computational overhead. Wang et al. [32] designed a fine-tuning-based FSOD framework that aligns visual features with class names and replaces linear classifiers with semantically similar classifiers. Multi-modal feature mixing is introduced to augment visual language communication so that trained similar base classes can explicitly support novel classes, and a maximum marginal loss of semantic perception is proposed to prevent class confusion. Han et al. proposed that VFA [17] further improves the performance of Meta-R-CNN by introducing variational feature learning into it. Wang et al. proposed that FPD [18] improves model performance by refining fine-grained prototypes to utilize the relationships between features.

Because of its effectiveness, the two-stage detection framework has been widely adopted in meta-learning-based FSOD methods. In the base training stage, many base class samples are used to train the model. In the fine-tuning stage, the model is fine-tuned with only K samples for each base and novel class. Meta-learning-based methods exhibit a high degree of flexibility and extensibility, enabling them to adapt easily to the introduction of novel classes. They can also be integrated with various meta-learning methods to improve performance further. These advantages make the meta-learning-based FSOD methods very effective in practical applications. Therefore, our method is based on the meta-learning method FPD, and we propose

to aggregate positive sample features and fuse negative sample features of the same class to improve the model's feature learning capability in the case of sample scarcity.

2.2 Contrastive Learning

Contrastive learning is a feature learning that aims to learn efficient representations by maximizing the similarity between same-class samples and the difference between different-class samples. The main idea is to train the model by comparing the relationships between different samples or between different views of the samples to extract differentiated feature representations. Contrastive learning can be categorized into unsupervised (UCL) and supervised (SCL) variants based on label dependency, differing in their utilization of annotated data [33].

2.2.1 Unsupervised Contrastive Learning

Unsupervised contrastive learning aims to learn meaningful representations from unlabeled data or classes. InstDisc [34] is an individual discriminant method that aims at self-supervised learning by treating each image as an independent class. This method uses a neural network to encode images into low-dimensional features. It optimizes these features to be separated as far as possible in the feature space, and negative samples are extracted from the Memory bank to augment feature differentiation. Based on the InstDisc, MoCo [35] replaces the Memory bank with a queue and proposes a momentum update encoder to improve the accuracy further. Simsiam [36] learns a more discriminative feature representation by minimizing the distance between predicted and real features.

2.2.2 Supervised Contrastive Learning

Supervised contrastive learning is a method of contrastive learning using labeled data. SupCon [37] generalizes the self-supervised batch contrastive paradigm to fully-supervised scenarios, enabling more effective utilization of label data. This paper optimizes the embedding space by attracting same-class clusters while repelling different clusters, enhancing inter-class separability and intra-class feature clustering. Chen et al. [38] incorporated a properly-weighted class-conditional InfoNCE loss and a class-conditional autoencoder into SupCon. The model performance is further improved. Contrastive Learning with Stronger Augmentations (CLSA) [39] uses the distribution between augmented images over the representation bank to supervise the retrieval of strongly augmented queries from the pool of instances.

Contrastive learning relies on substantial positive-negative sample pairs to enhance its performance in downstream tasks. We analyze existing contrastive learning-based object detection methods. CAReD [40] just proposed a contrastive learning network to supervise the training process, while VFA [17] introduced class-agnostic contrastive learning. However, since all support classes may be treated as negative, the contrastive effect is suboptimal. FPD [18] employed class-aware contrastive learning, but it only includes a single positive-negative sample pair, resulting in limited contrastive effectiveness. None of these methods addresses the sparsity of the feature space in few-shot scenarios. Our approach proposes generating mixed features to increase the amount of positive and negative sample pairs, alleviating feature space sparsity and enhancing model performance.

3 Method

In this section, we initially present the overall architecture of the task definition and model. Subsequently, we will delve into implicit contrastive learning and feature mixing, exploring their roles and implications in our framework.

3.1 Task Definition

In this work, we follow the standard few-shot object detection setting [8], utilizing two types of training datasets: base classes C_{base} with extensive information and novel classes C_{novel} with limited samples. The objective is to develop a detection model capable of recognizing novel objects during testing by effectively transferring information from the base classes.

3.2 The Model Architecture

Based on FPD, we propose an Implicit Feature Contrastive Learning method (IFCL), which consists of two parts: an implicit contrastive learning strategy and a mixed feature sampling method. By fusing the query feature and support feature, the implicit contrastive learning method implicitly learns the features of different-class samples so that the features of similar images are close in the feature space, and the features of dissimilar images are far away in the feature space. The mixed feature sampling method uses the mixup method to generate the mixed features for feature augmentation, which helps the model learn a more powerful feature representation.

As shown in Fig. 3, the framework employs a dual-branch siamese structure for joint query-support processing. First, the first three stages of the backbone network ResNet101 are used to extract the query features and support images. Then, the feature aggregation aggregates the support features into the query features. Next, we use the backbone network's fourth (final) stage to extract the high-level features of the two branches, which generate RoI features and support sample features, respectively. Finally, we use the IFCL module to generate mixed features and extract augmented positive and negative prototype features from the support sample features. These data are further processed by the feature fusion module and then fed into the detection head for final prediction.

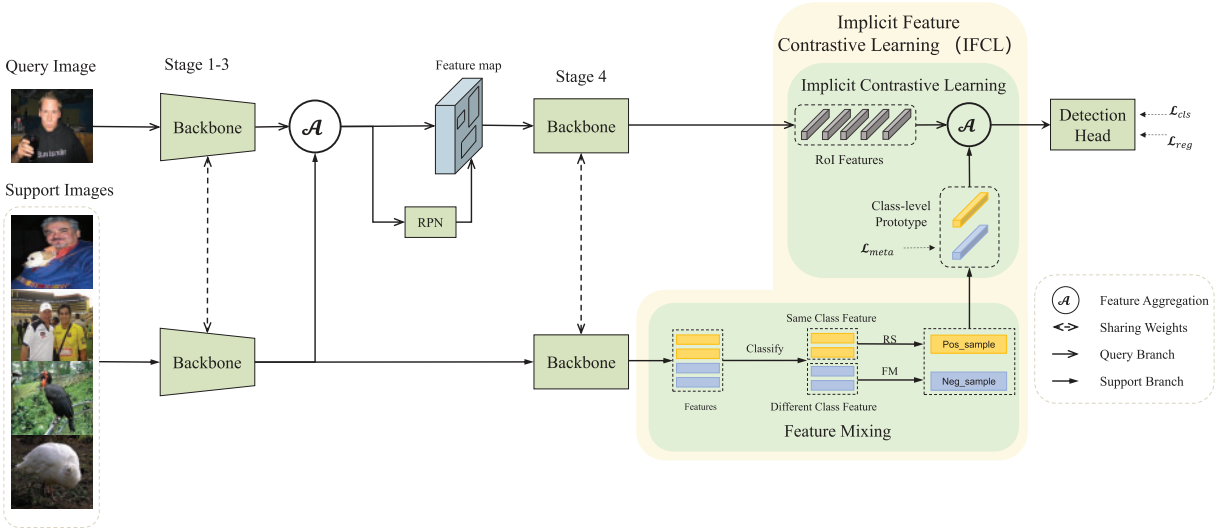

Figure 3: The overall architecture of our method (2-way 2-shot). IFCL is proposed to improve performance, whereas 'Classify' means grouping support samples based on annotation information to generate positive and negative samples. 'RS' means random selection to generate a positive sample, and 'FM' means feature mixing to generate a negative sample

3.3 Implicit Contrastive Learning

Inspired by contrastive learning, we analyze the feature aggregation method and find that it can be regarded as a contrastive learning strategy. The core idea of contrastive learning is to generate positive and negative sample pairs using data augmentation. These pairs are then compared, enabling the model to understand the relationships between samples better. This process also improves the model's generalization ability on unseen samples. The feature aggregation method, on the other hand, works by aggregating support sample features with query sample RoI features. By having the model learn the aggregated features, the intra-class and inter-class relationships are implicitly learned, thus improving the model's performance. This implicit contrastive mechanism makes feature aggregation not just a simple feature integration but a deep feature relationship learning.

With this implicit contrast, the model can learn the relationships between samples more efficiently in the feature space. This learning process improves the ability of the model to recognize samples of the same class. It also strengthens the ability of the model to distinguish between samples of different classes. In addition, this method belongs to the category of contrastive learning. It uses positive and negative sample features selected from the support features and aggregates them with the query features. This reflects an implicit property of contrastive learning.

3.4 Feature Mixing

In order to obtain augmented sample features, we first increased the amount of support samples for each epoch input model to carry out the same class sample mixing. As shown in Fig. 4, among the support samples, all the samples (f_1 and f_2) that are of the same class as the query sample f_q are extracted, and then one of them is randomly selected as the positive sample S_{pos}, while the negative sample S_{neg} is mixed by randomly selecting two samples (f_3 and f_4) from the remaining support samples with λ as the weight. Finally, a positive sample S_{pos} and a negative sample S_{neg} are obtained, which are subsequently aggregated in parallel with the RoI features. The mixing can be formulated as follows:

$$S_{neg} = \lambda * f_3 + (1 - \lambda) * f_4 \tag{1}$$

Considering that fusion of the average weight characteristics leads to dilution of important feature information (When $\alpha = \beta = 0.5$, the two features are mixed with equal weight. There is almost no change in the experimental result), we choose to use the beta distribution to generate weight coefficients as in the formula:

$$f(\lambda; \alpha; \beta) = \frac{\lambda^{\alpha-1} * (1 - \lambda)^{\beta-1}}{B(\alpha, \beta)} \tag{2}$$

In the base training stage, $S1$ and $S2$ are the same class as Q. $S3$ and $S4$ belong to the same class and are not the same class as Q. f_3 and f_4 are the features of the same class as shown in Fig. 4, which can better retain the consistency of the characteristics of this class and reduce the influence of noise from different classes. In the fine-tuning stage, only K samples ($K \leq 10$) for each class fine-tune the model. In order to increase the diversity and richness of features, f_2, f_3, and f_4 are the features of the different classes as shown in Fig. 5.

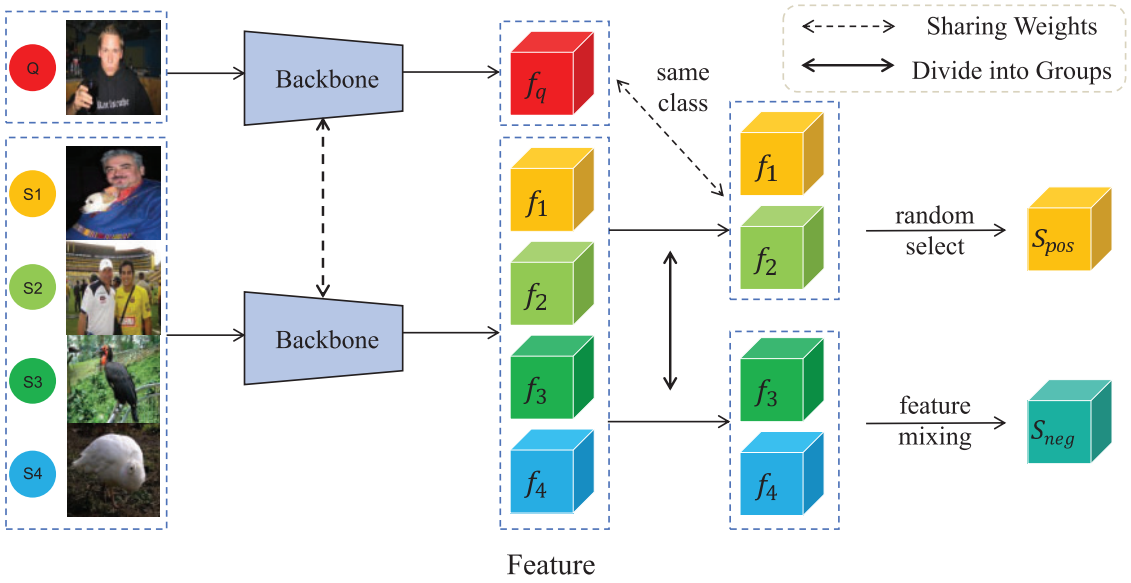

Figure 4: Based training mixed features sampling (2-way 2-shot). In the base training stage, S_{pos} is selected at random from two features with the same class, and S_{neg} is generated by mixing two features with the same class

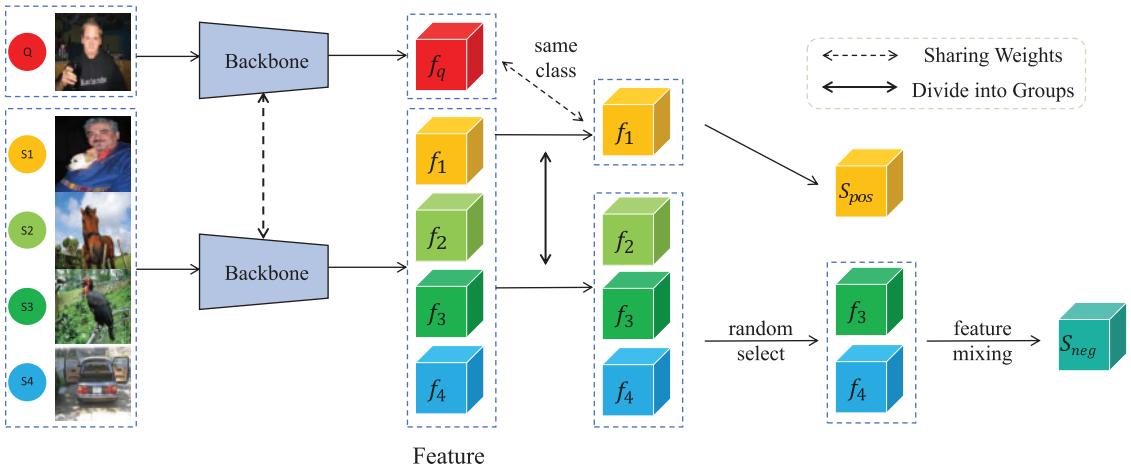

Figure 5: Fine-tuning mixed features sampling (4-way 1-shot). In base training stage, S_{neg} is generated by mixing two features selected from three different classes

4 Experiments

4.1 Datasets/Benchmark

We evaluate our method on a widely-used FSOD standard dataset, PASCAL VOC [41], using precisely the same class partitions and few-shot examples as in [5].

PASCAL VOC. We adopt the PASCAL VOC benchmark with 20 classes partitioned into 15 base classes and 5 novel classes. There are three different class splits for a more comprehensive evaluation. The model is trained on combined VOC2007 and VOC2012 train/val sets and evaluated on the VOC2007 test set. Performance is measured using IoU = 0.5 (mAP50) in multiple few-shot configurations $K = \{2, 3, 5, 10\}$ shot.

4.2 Implementation Details

Our method is implemented with MMDetection [42]. We adopt ResNet-101 [43] pretrained on ImageNet [44] as the backbone. The single scale feature map is used for detection without FPN [45].

All experiments are conducted using an NVIDIA RTX 3090 GPU, employing SGD optimization (momentum = 0.9), batch size = 4. The base training stage comprises 30,000 iterations on PASCAL VOC (initial learning rate = 0.005). During fine-tuning, we adopt a lower learning rate (learning rate = 0.001) while maintaining identical loss functions to Meta R-CNN. Evaluation metrics focus on novel class detection performance (nAP).

4.3 Comparison with the State-of-the-Art Methods

We present the results of a single experiment on the PASCAL VOC dataset in Table 1. It can be seen that IFCL is significantly better than previous methods, achieving state-of-the-art (SOTA) performance in most cases. Specifically, under the three Novel Sets of K = 10, IFCL is 3.2%, 1.8%, and 2.3% higher than the baseline, respectively.

Table 1: FSOD results on the PASCAL VOC dataset three Novel Sets (AP50)

Method/Shots	Backbone	Novel Set 1				Novel Set 2				Novel Set 3				Avg.
		2	3	5	10	2	3	5	10	2	3	5	10	
FSRW [8]	YOLOv2	15.5	26.7	33.9	47.2	15.3	22.7	30.1	40.5	25.6	28.4	42.8	45.9	31.18
MetaDet [15]	VGG16	20.6	30.2	36.8	49.6	23.1	27.8	31.7	43.0	23.9	29.4	43.9	44.1	33.64
Meta R-CNN [16]	ResNet-101	25.5	35.0	45.7	51.5	19.4	29.6	34.8	45.4	18.2	27.5	41.2	48.1	35.16
TFA w/cos [5]	ResNet-101	36.1	44.7	55.7	56.0	26.9	34.1	35.1	39.1	34.8	42.8	49.5	49.8	42.03
MPSR [20]	ResNet-101	42.5	51.4	55.2	61.8	29.3	39.2	39.9	47.8	41.8	42.3	48.0	49.7	45.74
Retentive [46]	ResNet-101	45.8	45.9	53.7	56.1	27.8	35.2	37.0	40.3	37.6	43.0	49.7	50.1	43.89
FSCE [47]	ResNet-101	43.8	51.4	61.9	63.4	29.5	43.5	44.2	50.2	41.9	47.5	54.6	58.5	49.19
Meta FR-CNN [30]	ResNet-101	54.5	**60.6**	**66.1**	65.4	35.5	46.1	47.8	51.4	46.4	53.4	59.9	58.6	53.85
Meta-DETR [31]	ResNet-101	51.4	58.0	59.2	63.6	36.6	43.7	49.1	54.6	45.9	52.7	58.9	60.6	52.86
FCT [48]	ResNet-101	57.1	57.9	63.2	67.1	34.5	43.7	49.2	51.2	**54.7**	52.3	57.0	58.7	53.88
VFA* [17]	ResNet-101	55.1	57.9	62.5	64.0	**42.5**	47.5	**50.7**	**52.0**	44.6	51.4	55.8	58.5	53.54
FPD* [18]	ResNet-101	52.8	58.1	63.9	64.3	39.9	46.4	49.1	50.1	45.2	52.5	58.6	59.0	53.33
ICFL(Ours)	ResNet-101	**58.6**	60.5	65.8	**67.5**	**42.5**	**47.7**	48.3	51.9	51.2	**54.4**	**60.3**	**61.3**	**56.67**

Note: '*' represents that the results are obtained by averaging over multiple replications. Bold and underlined indicate the best and the second-best results.

Furthermore, during the fine-tuning stage with a sample size of K = 2, our method demonstrates significant improvements in Novel Set 1 and Novel Set 3, outperforming the baseline FPD by 5.8% and 6.0%, respectively. This performance is highly competitive among comparable methods: compared to the meta-learning-based Meta-DETR, our method achieves substantial advantages across all three Novel Sets. Although Meta-DETR can rapidly adapt through task-level meta-training, its performance relies heavily on the distributional consistency between training and testing tasks. When faced with poor sample quality or significant inter-class feature variations, its generalization capability markedly declines, ultimately affecting detection accuracy. In contrast, our method employs a feature mixing approach that yet effectively expands the feature space, enhancing the model's discriminative ability. This makes the model learn robust feature representations even with an extreme lack of samples (K = 2). However, the performance improvement on Novel Set 2 is less pronounced compared to the other two Novel Sets. Although our method remains

comparable to VFA under the 2-shot condition, it does not exhibit a clear superiority. By analyzing the per-class accuracy of Novel Set 2 during fine-tuning (as shown in Table 2), we find that the detection accuracy of the "bottle" class significantly drags down the overall performance. As illustrated in Fig. 6, half of the "bottle" fine-tuning samples suffer from excessively small or incomplete objects, and the diversity in bottle shapes further complicates the ability of the model to learn effective features. Notably, this issue is more severe in Faster R-CNN-based methods. This is because the Meta Faster R-CNN-based method is highly sensitive to the target scale during the feature extraction stage. When confronted with small or incomplete objects, its Region Proposal Network (RPN) struggles to generate high-quality proposals, ultimately leading to suboptimal performance. In contrast, our method leverages implicit contrastive learning and feature mixing strategies. Fusing features across samples augments the feature representation capability for small objects, while the feature mixing implicitly improves the model's adaptability to shape diversity, thereby partially mitigating this issue.

Table 2: Results for each class on the PASCAL VOC dataset Novel Set 2 (AP50). The accuracy of the bottle class severely affects the overall accuracy

Class/Shot	Novel Set 2			
	2	**3**	**5**	**10**
Aeroplane	47.6	55.8	55.9	56.7
Bottle	9.1	11.6	12.7	19.0
Cow	56.9	54.7	55.8	62.0
Horse	62.6	64.9	61.9	65.0
Sofa	36.8	51.3	55.2	57.0
Avg.	42.5	43.0	48.3	51.9

Figure 6: Bottle class samples from Novel Set 2, where half of the general picture suffers from too small a scale or incomplete object

Under 5-shot and 10-shot conditions, IFCL underperforms compared to VFA. This is because VFA employs a class-agnostic feature aggregation strategy, treating all support samples as negative samples to

construct contrastive learning tasks. Although this design sacrifices some inter-class discriminability (e.g., Meta-DETR achieves 6% higher accuracy than VFA at K = 10), it effectively mitigates data bias issues. In contrast, our feature mixing strategy may introduce class bias. Overall, as shown in the "Avg." column of Table 1, our method achieves an average precision of 56.67%, surpassing existing state-of-the-art methods. The advantages in extremely few-shot scenarios, such as 2-shot and 3-shot settings, further validate the effectiveness of our implicit feature contrastive learning framework.

4.4 Analytical Experiment

4.4.1 Effect of Beta Distribution Parameters

In the process of generating augmented features by mixing samples, weight coefficients have a significant influence on the detection accuracy of the model. Beta distribution is a continuous probability distribution defined in the interval $[0, 1]$, λ defined by two shape parameters α and β that control the distribution's shape and degree of concentration. As in Fig. 7, when considered $\alpha = \beta, \lambda \approx 0.5$, it will lead to mixed sample features of equal weight coefficients, which dilutes important feature information. Moreover, at that time $|\alpha - \beta| \rightarrow \infty$, the weight coefficients $\lambda \rightarrow 1$, in this case, will form one feature as the main and another feature as a supplement, not only to avoid feature dilution but also to effectively improve the representation of features. The accuracy of (α, β) at different values, as shown in Fig. 8, when (α, β) takes the value of (12, 1), the detection accuracy is the highest. This result shows that using the weight coefficient λ generated in the case of (12, 1) can achieve the best results in increasing feature diversity and preventing the dilution of important features. In other cases, the detection accuracy is higher than the baseline, which proves the effectiveness of using Beta distribution to generate sample weights.

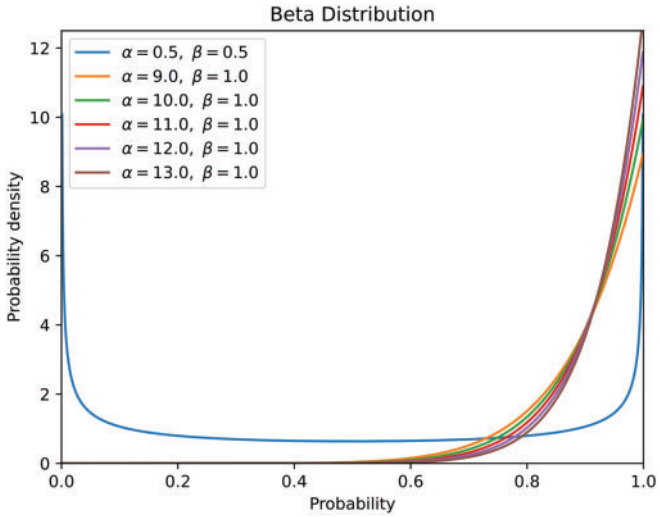

Figure 7: Beta distribution image, different curves have different probabilities of obtaining different weight coefficients. When $\alpha = \beta$, the value of λ always falls near 0.5

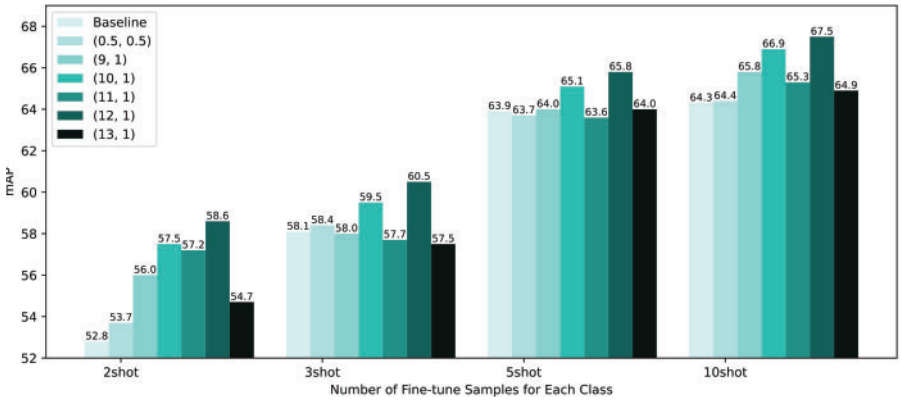

Figure 8: Analytical experiments on the selection of Beta distribution parameters, the average weight sample mixing results have basically no improvement in accuracy

4.4.2 Effect of Support Samples

In the base training stage, in order to mix the same class samples and, at the same time, ensure the condition that one feature is dominant and another feature is supplemented. We increased the amount of support samples input into the model for each iteration of base training from 15way 1shot to 15way 2shot, which ensures that the negative sample features are generated from a mixture of the same class sample features during the training process and achieve better results.

Also, because there is a sufficient amount of labeled data for each class in the base training process, we wondered if we could get better results if we increased the support samples to 15way Kshot ($2 < K < 5$) again and then randomly select two samples from K same class samples for feature mixing. Based on this idea, we do the following experiment, as shown in Fig. 9. The model detection accuracy is highest when the support sample is 15way 2shot. Increasing the support samples again leads to lower accuracy instead. We believe that increasing the amount of support samples again will increase the diversity of hybrid features. Since hybrid features are negative sample features, this feature diversity will interfere with the ability of the model to focus on the most critical positive samples, which will hurt performance.

In the fine-tuning stage, there are 20 classes of support samples, of which 15 classes are base classes, 5 are novel classes, and only N samples ($N \leq 10$) have labeled information in each class. In order to mix the same class of samples, we performed the same operation as the base training, increasing the 20-way 1-shot to the 20-way 2-shot. The result is shown in Fig. 10. The results of mixing the same class samples to generate the negative sample feature in the fine-tuning stage are worse. We analyze that this is due to insufficient samples in the dataset in the fine-tuning stage. Mixing the same class samples instead reduces the diversity of the features, which consequently influences the model's performance.

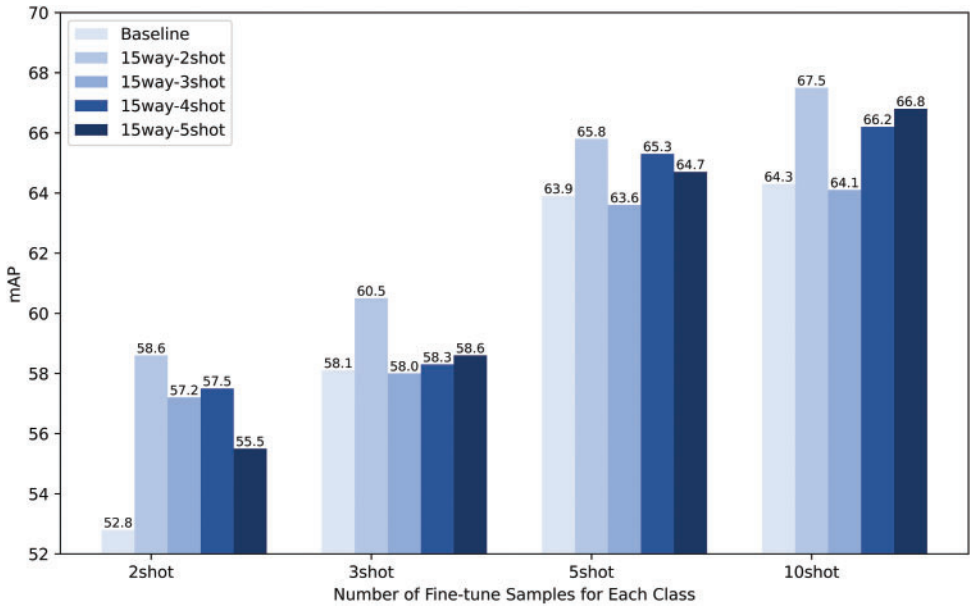

Figure 9: Analytical experiment of different amounts of samples for each class

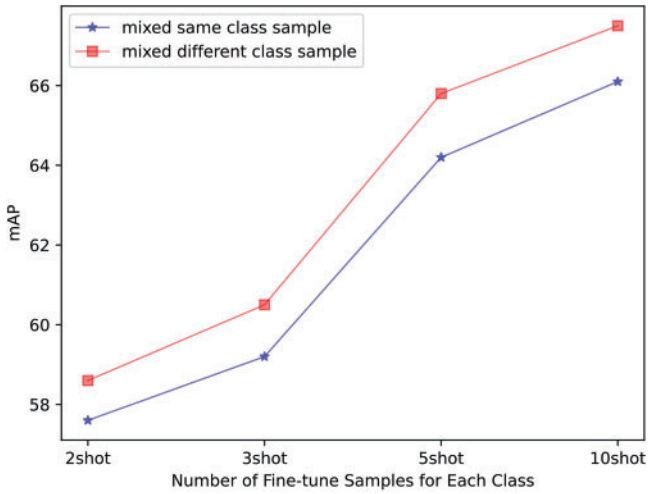

Figure 10: Analytical experiment of different amounts of samples for each class. '⋆' means mixed same class sample feature, '□' means mixed different class sample feature

4.4.3 Computational Efficiency Analysis

To evaluate the practicality of our approach, we contrast IFCL with other approaches in terms of computational time during the training and testing stage (using data from Novel Set 1). All experiments are conducted under identical hardware (NVIDIA RTX 3090 GPU) and software (PyTorch 1.12.0) environments to ensure fairness. We record the training time for each method on the three data splits of the PASCAL VOC dataset, including both base-training and fine-tuning stages. As shown in Table 3, both VFA and FPD methods are improved versions of the Meta-R CNN framework. Our IFCL further improves the FPD method based on this framework. The results show that VFA and FPD have similar training and inference times, slightly longer than Meta R-CNN. Compared to FPD (0.40 and 0.38 s/epoch), IFCL requires

marginally more training time (0.46 and 0.39 s/epoch) due to the additional computational overhead from feature mixing and implicit contrastive learning. During inference, the average processing time per image for IFCL differs by no more than 7.2 ms/image compared to FPD. Although IFCL introduces additional computational costs in training and testing, its performance improvements in few-shot object detection tasks (as shown in Table 1) justify these costs. Furthermore, IFCL's computational efficiency will improve passively as computer hardware advances, while its algorithmic advantages in feature mixing will maintain long-term value.

Table 3: The inference time of the model during the training and testing stages

Methods	Novel Set 1		
	Base-training (s/e)	**Fine-tuning (s/e)**	**Test (s/i)**
Meta R-CNN	0.35	0.33	$5.35 * 10^{-2}$
VFA	0.39	0.35	$5.78 * 10^{-2}$
FPD	0.40	0.38	$5.82 * 10^{-2}$
IFCL(Ours)	0.46	0.39	$6.54 * 10^{-2}$

4.5 Ablation Experiment

Compared with the FPD approach, we propose an implicit feature contrastive learning module that improves the ability of the model to learn feature representations by generating augmented features. To verify the effect of this module on the experimental, we conduct relevant ablation experiments for the base training and fine-tuning stage on Novel Set 1, and the results are listed in Table 4. It can be seen that the effect of applying the IFCL module only in the base training stage is higher than that of applying the IFCL module only in the fine-tuning stage. This is because the goal of the base training stage is to let the model learn the basic feature representation. At this stage, the introduction of the IFCL module can help the model better capture important features, improve the feature discrimination ability by using a lot of training data, promote the optimization of the model in an effective direction, and enhance the diversity and robustness of features. Make the model perform better when learning different class boundaries, thereby improving overall performance. In contrast, in the fine-tuning stage, the limited annotation data makes the feature representation learned by the model less targeted, leading to poor results. In general, the use of the IFCL module in both stages has played a positive role in the detection of small sample targets.

Table 4: Ablation experiment on implicit feature contrastive learning module

Methods/Shots	IFCL		Novel Set 1			
	BT	**FT**	**2**	**3**	**5**	**10**
Baseline	×	×	52.8	58.1	63.9	64.3
Ours	×	✓	54.3	58.8	64.4	64.7
Ours	✓	×	58.2	60.0	65.4	65.9
Ours	✓	✓	**58.6**	**60.5**	**65.8**	**67.5**

Note: 'BT' means the Base-Training stage, and 'FT' means the Fine-Tuning stage.

5 Conclusion

Aiming at the problem of sample scarcity in few-shot learning in order to help the model learn a more powerful boosted representation, we propose a new concept of implicit contrastive learning by analyzing the connection between contrastive learning and feature aggregation methods and further propose an implicit feature contrastive learning module introducing hybrid feature sampling for generating augmented features. The sound performance on the PASCAL VOC dataset proves the effectiveness of the implicit feature comparison learning module. In addition, in the analysis experiments, we find that completely average mix features will dilute the important feature information, which is not conducive to the model learning feature representation. In contrast, introducing a small amount of other sample features into a sample feature can improve the feature diversity while retaining the important feature information, which can effectively help the model learn useful feature representations.

Despite the results achieved in this study, there are still shortcomings. Firstly, we introduced the feature mixing method, which improved the model complexity and increased the training time; in the future, we may consider designing a more lightweight model structure to reduce the model complexity. Secondly, there is still room for improvement in the model's ability to recognize classes where some sample objects are incomplete or too small, and the introduction of a more complex implicit contrastive learning strategy may be considered in the future. In addition, the experiment only uses the PASCAL VOC dataset, and future work will extend to more domains and datasets for validation. Finally, the hyper-parameter settings for the beta distribution for implicit feature contrastive learning and their impact on model performance have not been explored in depth in this study. Future research will focus on optimizing to improve the model's performance further.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This study was funded by the China Chongqing Municipal Science and Technology Bureau, grant numbers CSTB2024TIAD-CYKJCXX0009, CSTB2024NSCQ-LZX0043, CSTB2022NSCQ-MSX0288; Chongqing Municipal Commission of Housing and Urban-Rural Development, grant number CKZ2024-87; the Chongqing University of Technology Graduate Education High-Quality Development Project, grant number gzlsz202401; the Chongqing University of Technology—Chongqing LINGLUE Technology Co., Ltd. Electronic Information (Artificial Intelligence) Graduate Joint Training Base; the Postgraduate Education and Teaching Reform Research Project in Chongqing, grant number yjg213116; and the Chongqing University of Technology-CISDI Chongqing Information Technology Co., Ltd. Computer Technology Graduate Joint Training Base.

Author Contributions: Gang Li: Methodology, Writing—original draft, Writing—review & editing, Funding acquisition. Zheng Zhou: Writing—original draft, Methodology, Investigation. Yang Zhang: Methodology, Writing—review & editing, Funding acquisition, Project administration. Zihan Ruan: Survey, Research. Ru Wang: Formal analysis, Visualization. Xinyu Fan: Validation, Formal analysis. Pengfei Lv: Experimental verification. Wei Tan: Visualization. Chuanyun Xu: Writing—review & editing, Supervision, Resources, Funding acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All relevant data are within the paper. The data are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. p. 213–29.
2. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. p. 21–37.
3. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88.
4. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transact Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
5. Wang X, Huang TE, Darrell T, Gonzalez JE, Yu F. Frustratingly simple few-shot object detection. In: Proceedings of the 37th International Conference on Machine Learning, ICML'20; 2020 Jul 13–18; New York, NY, USA; Online. p. 9919–28.
6. Cao Y, Wang J, Jin Y, Wu T, Chen K, Liu Z, et al. Few-shot object detection via association and discrimination. Adv Neural Inform Process Syst. 2021;34:16570–81.
7. Qiao L, Zhao Y, Li Z, Qiu X, Wu J, Zhang C. DeFRCN: decoupled faster R-CNN for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 8681–90.
8. Kang B, Liu Z, Wang X, Yu F, Feng J, Darrell T. Few-shot object detection via feature reweighting. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8420–9.
9. Sun Q, Liu Y, Chua TS, Schiele B. Meta-transfer learning for few-shot learning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 403–12.
10. Wu Z, Li H, Zhang D. Few-shot object detection via transfer learning and contrastive reweighting. In: International Conference on Artificial Neural Networks; 2023 Sep 26–29; Crete, Greece. p. 78–87.
11. Gao BB, Chen X, Huang Z, Nie C, Liu J, Lai J, et al. Decoupling classifier for boosting few-shot object detection and instance segmentation. Adv Neural Inform Process Syst. 2022;35:18640–52.
12. Shangguan Z, Huai L, Liu T, Liu Y, Jiang X. Decoupled DETR for few-shot object detection. In: Proceedings of the 2024 Asian Conference on Computer Vision (ACCV); 2024 Dec 8–12; Hanoi, Vietnam. p. 286–302.
13. Demirel B, Baran OB, Cinbis RG. Meta-tuning loss functions and data augmentation for few-shot object detection. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 22; Vancouver, BC, Canada. p. 7339–49.
14. Fang H, Han B, Zhang S, Zhou S, Hu C, Ye WM. Data augmentation for object detection via controllable diffusion models. In: Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision; 2024 Jan 3–8; Waikoloa, HI, USA. p. 1257–66.
15. Wang YX, Ramanan D, Hebert M. Meta-learning to detect rare objects. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9925–34.
16. Yan X, Chen Z, Xu A, Wang X, Liang X, Lin L. Meta R-CNN: towards general solver for instance-level low-shot learning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9577–86.
17. Han J, Ren Y, Ding J, Yan K, Xia GS. Few-shot object detection via variational feature aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2023 Feb 7–14; Washington, DC, USA. p. 755–63.
18. Wang Z, Yang B, Yue H, Ma Z. Fine-grained prototypes distillation for few-shot object detection. In: Proceedings of the 2024 AAAI Conference on Artificial Intelligence; 2024 Feb 26–27; Vancouver, BC, Canada. p. 5859–66.
19. Köhler M, Eisenbach M, Gross HM. Few-shot object detection: a comprehensive survey. IEEE Transact on Neural Netw Learn Syst. 2024;35(9):11958–78. doi:10.1109/tnnls.2023.3265051.
20. Wu J, Liu S, Huang D, Wang Y. Multi-scale positive sample refinement for few-shot object detection. In: Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 456–72.

21. Zhang W, Wang YX. Hallucination improves few-shot object detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 13008–17.
22. Riou K, Zhu J, Ling S, Piquet M, Truffault V, Le Callet P. Few-shot object detection in real life: case study on auto-harvest. In: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP); 2020 Sep 21–23; Online. p. 1–6.
23. Zhu C, Chen F, Ahmed U, Shen Z, Savvides M. Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 8782–91.
24. Rahman S, Khan S, Barnes N, Khan FS. Any-shot object detection. In: Proceedings of the 2020 Asian Conference on Computer Vision; 2020 Nov 30–Dec 4; Kyoto, Japan. p. 89–106.
25. Chen Z, Fu Y, Zhang Y, Jiang YG, Xue X, Sigal L. Multi-level semantic feature augmentation for one-shot learning. IEEE Transact Image Process. 2019;28(9):4594–605. doi:10.1109/tip.2019.2910052.
26. Huang L, Dai S, He Z. Few-shot object detection with semantic enhancement and semantic prototype contrastive learning. Knowl Based Syst. 2022;252(4):109411. doi:10.1016/j.knosys.2022.109411.
27. Huang X, Choi SH. Feature-semantic augmentation network for few-shot open-set recognition. Pattern Recognit. 2024;156(8):110781. doi:10.1016/j.patcog.2024.110781.
28. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3339–48.
29. Chen H, Wang Y, Wang G, Qiao Y. LSTD: a low-shot transfer detector for object detection. In: Proceedings of the 2018 AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA. p. 2836–43.
30. Han G, Huang S, Ma J, He Y, Chang SF. Meta faster R-CNN: towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Feb 22–Mar 1; Menlo Park, CA, USA; Online. p. 780–9.
31. Zhang G, Luo Z, Cui K, Lu S, Xing EP. Meta-DETR: image-level few-shot detection with inter-class correlation exploitation. IEEE Transact Pattern Anal Mach Intellig. 2022;45(11):12832–43. doi:10.1109/TPAMI.2022.3195735.
32. Wang Z, Gao Y, Liu Q, Wang Y. Semantic enhanced few-shot object detection. In: 2024 IEEE International Conference on Image Processing (ICIP); 2024 Oct 27–30; Abu Dhabi, UAE. p. 575–81.
33. Chen J, Qin D, Hou D, Zhang J, Deng M, Sun G. Multiscale object contrastive learning-derived few-shot object detection in VHR imagery. IEEE Transact Geosci Remote Sen. 2022;60:1–15. doi:10.1109/tgrs.2022.3229041.
34. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3733–42.
35. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 9729–38.
36. Chen X, He K. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 15750–8.
37. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. Adv Neural Inform Process Syst. 2020;33:18661–73.
38. Chen M, Fu DY, Narayan A, Zhang M, Song Z, Fatahalian K, et al. Perfectly balanced: improving transfer and robustness of supervised contrastive learning. In: International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 3090–122.
39. Wang X, Qi GJ. Contrastive learning with stronger augmentations. IEEE Transact Pattern Analys Mach Intell. 2022;45(5):5549–60. doi:10.1109/tpami.2022.3203630.
40. Quan J, Ge B, Chen L. Cross attention redistribution with contrastive learning for few shot object detection. Displays. 2022;72(2):102162. doi:10.1016/j.displa.2022.102162.
41. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Computer Vis. 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.

42. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, et al. MMDetection: open mmlab detection toolbox and benchmark. arXiv:1906.07155. 2019.
43. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Computer Vis. 2015;115(3):211–52. doi:10.1007/s11263-015-0816-y.
45. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 2117–25.
46. Fan Z, Ma Y, Li Z, Sun J. Generalized few-shot object detection without forgetting. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 4527–36.
47. Sun B, Li B, Cai S, Yuan Y, Zhang C. FSCE: few-shot object detection via contrastive proposal encoding. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 7352–62.
48. Han G, Ma J, Huang S, Chen L, Chang SF. Few-shot object detection with fully cross-transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5321–30.