Few-Shot Object Detection via Instance-wise and Prototypical Contrastive Learning

Qiaoning Lei, Yinsai Guo, Liyan Ma, Xiangfeng Luo*
School of Computer Engineering and Science, Shanghai University
Shanghai, China
Email: luoxf@shu.edu.cn

***Abstract*—Few-shot object detection (FSOD), which involves training the detector with few annotated data to detect novel objects, has aroused a wide range of research interests. However, the performance of FSOD is still limited by insufficient data. Existing works usually adopt fine-tuning paradigm, which first uses rich base classes for pre-training and then uses them to carve the novel class feature space. In the fine-tuning phase, the balance space learned by the pre-trained model will be broken leading to an intersection between the feature space of novel and base classes, which makes it difficult to distinguish the difference between them. Contrastive learning has been shown to learn a balanced feature space and enhance the discriminability of the learned features. Here, we present Few-Shot object detection via Instance-wise and Prototype Contrastive Learning (FS-IPCL), which introduces contrastive learning to learn a balanced feature space. FS-IPCL uses instance-wise and prototype contrastive loss during feature learning to enhance the intra-class compactness and inter-class separability of samples. In this way, the base and novel classes can be evenly distributed in the feature space, improving the class boundary to alleviate the confusion problem of the novel classes. Extensive experimental results on the PASCAL VOC and MS-COCO datasets demonstrate the effectiveness of the proposed method and achieve state-of-the-art performance.**

***Index Terms*—Few-shot object detection; Contrastive learning; Graph convolutional network.**

I. Introduction

With the widespread use of Deep Convolutional Neural Networks (DCNN) [1], [2] in recent years, object detection [3] algorithms based on DCNN have also advanced significantly. To assure the effectiveness of its detectors, object detection nevertheless depends on a sizable volume of annotated data. However, it requires considerable labor and time to get enough annotated data in the real world. Deep detectors are prone to overfitting when trained with only a few data, and their detection accuracy cannot be compared with that of detectors with a large amount of data. As opposed to that a child may quickly pick up novel categories and visual concepts by using a few examples. Therefore, most researchers are working to close this gap to give machines good perception capabilities.

Few-shot learning follows the principle of picking up new ideas rapidly, which can achieve promising performance with limited data. As a branch of few-shot learning, few-shot object detection (FSOD) is a more complex task compared to few-shot classification because it requires to locate the objects additional. The majority of FSOD approaches follow the meta-learning paradigm. The two-stage fine-tuning method (TFA) [4] has recently demonstrated more potential for improving FSOD because of its efficiency and simplicity compared to meta-learning techniques. To address the issue of sparse scale distribution of objects in FSOD, MPSR [5] proposes a multi-scale positive sample refining method on the basis of TFA. Nevertheless, the application difficulty of the method is increased due to the requirement for manual selection in its forward refinement branch. In this work, we find that fine-tuning using novel class samples is hard to learn a balanced discriminative feature space, making it difficult for the model to distinguish the features of the base classes and the novel classes.

Contrastive learning has been shown to learn a balanced feature space [6]. Since the distances and differences between categories are more obvious, contrastive learning can enhance the discriminability of the learned features. FSCE [7] uses supervised batch contrastive learning [8] to simulate the instance-level similarity and inter-class distinction of object proposal embeddings. It produces a more balanced feature space by optimizing the instance contrast loss, which can separate all instances well. However, since the model can distinguish different instances using low-level image differences, the learned feature embeddings may not capture the semantic knowledge of the objects. Inspired by FSCE, we introduce contrastive learning into FSOD with prototype contrastive learning. Prototypes are described as “representative embeddings of a class of semantically similar instances” [9]. It aggregates instances of the same class to obtain a compact image representation, which can capture some basic semantic structure of a single class.

We propose Few-Shot object detection via Instance-wise and Prototypical Contrastive Learning (FS-IPCL). Specifically, the instance-wise contrastive loss (ICL) improves the similarity of instances from the same class, and the prototype contrastive loss (PCL) improves the similarity between an instance and its corresponding prototype, so that the model learns more class-related semantic knowledge. However, two problems arise when contrastive learning is applied to FSOD: (1) there may be background noise in contrast samples; (2) there may exist positive and negative sample coupling [10]. That is, low quality positive samples will decrease the gradient

*Corresponding author.

DOI reference number: 10.18293/SEKE2023-129

of a batch of informative negative samples and vice versa. To address these two problems, we screen contrast samples to obtain high-quality positive and negative samples and decoupling them. Additionally, the prototype constructed using the features extracted in the initial stage may not be very reliable, and the feature representation of the image will constantly change during training. To obtain a representative prototype embedding, we adopt graph convolutional network (GCN) [11] to dynamically update the prototypes. Our approach has three main contributions:

- The instance and prototype contrastive loss are designed to learn a balanced discriminative feature space, improving by selecting high-quality positive and negative samples and decoupling them.
- To get more accurate prototype feature embeddings, GCN is used to dynamically update the prototypes in prototypical contrastive loss.
- A significant number of experiments are conducted by our method on the Pascal VOC and MS-COCO datasets, and new state-of-the-art results are achieved.

II. Relation Work

Few-Shot Object Detection. Currently, the commonly used FSOD methods are mainly based on meta-learning and fine-tuning paradigms. Following the meta learning methods, FSRW [12] proposes a reweighting module to extract the global features of the support image. For fine-tuning paradigm, TFA [4] first applies the two-stage fine-tuning method to FSOD, and proposes a new few-shot evaluation method. DeFRCN [13] uses a decoupling approach to solve the multi-stage and multi-task conflict problem when the Faster R-CNN detection framework is applied to FSOD. Our approach is also based on a two-stage fine-tuning paradigm.

Contrastive learning. Contrastive learning uses contrast to bring together similar classes and distinguish classes by constructing pairs of positive and negative samples. Most contrastive learning methods map features to unit hyperspheres for representation learning. A direct matching of uniformly sampled points on the unit hypersphere can provide a good representation. Existing contrastive learning is divided into two main categories, instance-wise contrastive learning [6] and prototypical contrastive learning [9]. Instance-wise contrastive learning brings similar instances close together and different instances far apart by comparing instance information from different samples. Prototypical contrastive learning represents the semantic structure of a class by using prototypes of image clusters. Some methods [14] have also used both instance-wise contrast learning and prototypical contrast learning. In this paper, contrastive learning is introduced into FSOD by mapping features onto unit hypersphere and instance-wise contrast learning and prototype contrast learning.

III. Method

Our proposed method FS-IPCL uses two-stage training. The base classes training step uses a rich base-class dataset to train the model. During the fine-tuning phase, we fine-tune the model using a relatively balanced small amount of base-class data and novel classes data (N-way K-shot). And the backbone feature extractor is frozen while the rest of the structure is fine-tuned. Meanwhile, we introduce ICL and PCL to supervise the RoI feature extractor, and jointly optimize the contrastive loss and the original classification and regression objectives. Our method's overall architecture is shown in Figure 1.

A. Problem Setting

Our problem setting for FSOD follows the standard problem setting of previous works [4], [7], [13]. Our dataset is divided into a base set D_{base} with rich annotated instances, and a novel support set D_{novel} with only a few annotated instances per category. The class C_{base} in our base set D_{base} and the class C_{novel} in the novel support set D_{novel} do not overlap, that is, $C_{base} \cap C_{novel} = \varnothing$. Our goal is to learn a robust detector which can recognize and localize the query set D_{query} pairs without annotated instances, where the class $C_{query} \subseteq C_{base} \cup C_{novel}$ in D_{query}.

B. Instance-wise and Prototypical Contrastive Loss

Since generic detectors struggle to capture discriminative region proposal features from a limited number of shots, we propose two contrastive losses to better distinguish feature representations of the novel classes: (1) ICL reduces similarity between object candidates from different classes and increases similarity between object candidates within the same category by comparing various RoI features. (2) PCL compares the RoI feature of the object with the prototype, making the object candidate box close to its corresponding class prototype and far away from other class prototypes. PCL first acquires the initial prototype $c^{'} \in \mathbb{R}^{d_p \times K}$ and then utilizes the prototype updating operation to dynamically update the prototype. The prototype updates part which is described in III-C.

Inspired by supervised batch contrastive learning method [8], our ICL and PCL are designed as follows with suitable for FSOD. Specifically, we adopt the selected N^+ positive samples and N^- negative samples to reduce the noise influence in the samples. And we remove the similarity calculation of positive samples in the denominator to alleviate the influence of the coupling of positive and negative samples.

$$L_{ICL} = \frac{-1}{N^+} \sum_{i=1}^{N^+} \frac{1}{N_{yi}-1} \sum_{j=1,\ j\neq i}^{N^+} \mathbb{I}\left\{y_i = y_j\right\} \cdot log \frac{exp\left(\widetilde{z}_i \cdot \widetilde{z}_j/\tau\right)}{\sum_{k=1}^{N^+ \cup N^-} \mathbb{I}\left\{y_i \neq y_k\right\} \cdot exp\left(\widetilde{z}_i \cdot \widetilde{z}_k/\tau\right)}, \tag{1}$$

$$L_{PCL} = \frac{-1}{N^+} \sum_{i=1}^{N^+} \frac{1}{N_{yi}-1} \sum_{j=1}^{K} \mathbb{I}\left\{y_i = y_j\right\} \cdot log \frac{exp\left(\widetilde{z}_i \cdot \widetilde{c}_j/\tau\right)}{\sum_{k=1}^{K} \mathbb{I}\left\{y_i \neq y_k\right\} \cdot exp\left(\widetilde{z}_i \cdot \widetilde{c}_k/\tau\right)}, \tag{2}$$

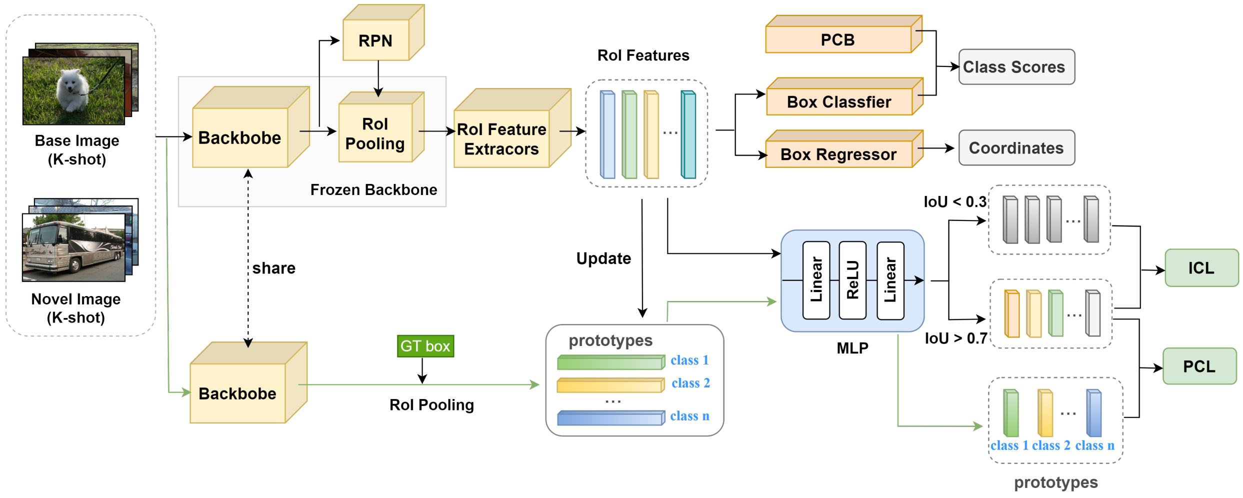

Fig. 1. Overview of our proposed FS-IPCL. The RPN and RoI feature extractors are fine-tuned in our approach in addition to the bounding box classifier and bounding box regressor. To introduce contrast learning, we select the RoI features for ICL, extracting features for the ground truth of all support images to compute prototype vectors for each class, and utilize them for PCL, while dynamically updating the prototypes. The intra-class consistency and inter-class separability are maximized by optimizing ICL and PCL.

where y_i is the label of the ground truth, and $\widetilde{z}_i = \frac{z_i}{\|z_i\|}$ represents the feature normalization operation. $\widetilde{z}_i \cdot \widetilde{z}_j$ denotes the inner (dot) product between the $i-th$ and $j-th$ proposal in the projected hypersphere denotes the inner (dot) product. $\widetilde{z}_i \cdot \widetilde{c}_j$ denotes the inner product between the $i-th$ proposal and the $j-th$ class prototype feature. τ denotes the temperature parameter.

Sample selection strategy. Unlike image classification, which uses the semantic information of the whole image, our model classifies based on the classification signals in the candidate boxes generated by RPN. Considering that most of the candidate boxes may be offset from the object instance, the information in the candidate box does not describe the corresponding instance accurately enough, so the contrast samples are screened. We use the Intersection over Union (IoU) score u_i between proposal and its matching ground truth bounding box to select the corresponding samples. Referring to most of the IoU threshold screening methods, we take the samples with u_i greater than 0.7 as the foreground instances of the N^+ samples. At the same time, N^- samples with u_i less than 0.3 are selected as background samples. By selecting the samples, we can obtain the instance objects containing more information and reduce the background interference in the comparison. Besides, the selecting eliminates most of the samples and reduces the amount of calculation of the model.

Decoupling strategy for positive and negative samples. Inspired by decoupled contrastive learning [10], we further modify the contrastive loss. Decoupled contrastive learning uses a large number of experiments to prove that the currently widely used cross-entropy (InfoNCE) [15] loss has obvious positive and negative coupling effects, which reduces the effect of the model in small batch learning and affects the training efficiency of the model. Therefore, we use the decoupling strategy of positive and negative samples in the contrast loss. By directly removing the similarity between the positive sample pairs in the denominator, the ratio of the sum of similarities for all positive sample pairs to the sum of similarities for all negative sample pairs can be computed directly. This allows the model to optimize positive and negative sample pairs separately, decoupling their influences. And the training efficiency of contrastive learning is improved.

Subsequently, we use the Prototype Calibration Module (PCB) in [13] to further refine our classification scores. And the classifier of cosine similarity in TFA is adopted in the fine-tuning step.

C. Prototype Updates

Since the representation of the image is constantly updated during the training process, the prototype feature embeddings need to be maintained to make the obtained prototype-like feature embeddings more representative. Therefore, we dynamically update the prototypes during training. The prototype update part is shown in Figure 2. For K classes, their initial prototypes are first obtained. The backbone network is used to extract original image features for a given support set S, and then RoI Pooling and ground-truth box are used to obtain the object instance embedding representation f_i. The initial prototype representation of each class is obtained by means of the mean value. The formula is:

$$c_K^{'} = \frac{1}{|S_K|} \sum_{(f_i,y_i)\in S_K} f_i. \tag{3}$$

Since the candidate boxes may be biased and contain a lot of background noise, we screen the N features used to update. The B RoI features are selected by the IoU score u_i of the prediction box and the ground truth of the corresponding prediction class greater than threshold φ.

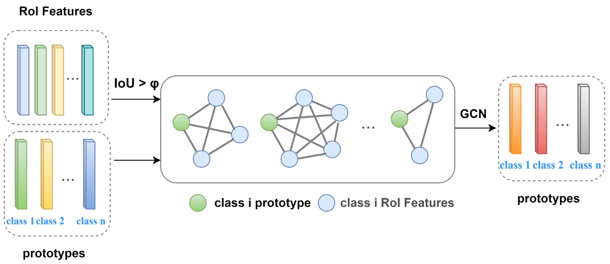

Fig. 2. Each subgraph consists of a class prototype feature and its corresponding class RoI features as nodes, and the edge is the cosine similarity between the feature vectors. The information of RoI features is aggregated into the corresponding class prototypes by the way of GCN.

K subgraphs $G = (V, E)$ are constructed by using the class prototypes and the selected B RoI features, where V and E represent the node set and edge set. The node set V of each subgraph consists of a class prototype and the RoI features of its corresponding class. Edges are constructed between prototype and RoI features, and between RoI features to represent the relationships between them. Thereinto, the information from the RoI features of the associated class can be transferred to the prototype to update its features by creating edges between the prototype and RoI features. By constructing edges between RoI features, different RoI feature information can be used to enrich the original RoI feature information. The adjacency matrix A is used to represent this relationship. In general, considering that the same class have more similar features, we adopt cosine similarity to obtain the similarity between prototype and RoI features, and between RoI features, thereby obtaining the adjacency matrix A. Then, it is normalized to acquire the normalized adjacency matrix $A^{'}$.The formula is:

$$A_K^{x_i^K} = \frac{\left(c_K^{'}\right)^{\top} x_i^K}{\left\|c_K^{'}\right\|_2 \cdot \left\|x_i^K\right\|_2}, \tag{4}$$

$$A^{'} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{5}$$

where D denotes the diagonal degree matrix and $D_{ii} = \sum_j A_{ij}$. For simplicity, $A_K^{x_i^K}$ is simply denoted as A, and the similarity calculation between RoI features is also calculated by referring to (4).

Using the feature similarity size provided by the adjacency matrix A, we adopt GCN to aggregate the feature information of the candidate boxes into the corresponding class prototypes. For each prototype c_K , the effect of each layer of GCN is equivalent to the weight sum of its corresponding class's RoI features, and each prototype is updated as follows:

$$c_K = \frac{1}{B_K} c_K^{'} + \sum_{x_i^K \in x_{B_K}} A^{'} \cdot x_i^K. \tag{6}$$

D. Loss Function

The loss function consists of the classification and bounding box loss functions of RPN and RCNN in Faster R-CNN, as well as the ICL and PCL of the contrastive learning module.

$$L = L_{rpn} + L_{cls} + L_{reg} + \lambda_{icl} L_{ICL} + \lambda_{pcl} L_{PCL}, \tag{7}$$

λ_{icl} and λ_{pcl} are set to 0.5 respectively, to balance the loss.

IV. Experiments

In this section, we first provide a description of the few-shot object detection datasets and detection settings. Following this, we will present our detection results and ablation studies of our approach on Pascal VOC dataset. Finally, we provide the detection results on the MS COCO benchmark.

A. Experimental Setting

Existing benchmarks. We adopt the dataset settings of previous works [4], [13] to ensure that our method can be fairly compared. As for PASCAL VOC, we use three random split groups with 20 classes each, and each group is randomly classified into 15 base classes and 5 novel classes. Each novel class contains K=1/2/3/5/10 annotated samples from the combination of PASCAL VOC's 2007 and 2012's trainval. The detection ability of novel classes is assessed using nAP50, which is the IoU threshold with an average precision of 0.5 for the novel classes. For MS-COCO, we split the 80 classes into two separate datasets: a base class dataset consisting of 60 classes, and a novel class dataset consisting of the remaining 20 classes, where K=10/30. Similarly, we evaluate the detection performance of novel classes by employing nAP and nAP75 with different IoU thresholds.

Implementation Details. Our model framework uses Faster R-CNN [3], and ResNet-101 [1] is adapted as our backbone. All experiments are trained on 1 RTX3090 GPU, and the batch size is set to 4. The solver using standard SGD with momentum set to 0.09 and weight decay of 5e-5. During the base training phase, we set the learning rate to 0.005, and adjust it to 0.01 in the fine-tuning phase.

B. Experiments on PASCAL VOC

1) Comparisons with State-of-the-art Methods: Our method is compared with the previous state-of-the-art methods on the experimental results of three random splits of PASCAL VOC, as shown in Table I. It can be seen that FS-IPCL demonstrates better performance than existing FSOD methods in the majority of experimental conditions. Specifically, for Novel Split 1, our method improves by 3.8% (40.2% vs. 44.0%) in the 1-shot setting, 3.6% (53.6% vs. 57.2%) in the 2-shot setting, an improvement of 0.8% (58.2% vs. 59.0%) in the 3-shot setting and 0.7% (63.6% vs. 64.3%) in the 5-shot setting. In the 10-shot scenario, our method is on par with the current best performing method and a 0.8% improvement over the second best performing method (59.7% vs. 66.5%). In the remaining two splits, although the existence of FS-IPCL is lower than other methods, the

TABLE I

Experimental results on three splits of Pascal VOC. We evaluate the performance of our method on 3 splits using nAP50. Boldface indicates SOTA results.

Method / Shots	Venue	Novel Split 1					Novel Split 2					Novel Split 3					Avg.
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FRCN-ft [16]	ICCV 2019	9.9	15.6	31.6	38.0	52.0	9.4	13.8	17.4	21.9	39.7	8.1	13.9	19	23.9	44.6	24.3
FSRW [12]	ICCV 2019	14.2	23.6	29.8	36.5	35.6	12.3	19.6	25.1	31.4	29.8	12.5	21.3	26.8	33.8	31.0	25.6
TFA w/cos [4]	ICML 2020	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6	34.7
Viewpoint [17]	ECCV 2020	24.2	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
MPSR [5]	ECCV 2020	34.7	42.6	46.1	49.4	56.7	22.6	30.5	31.0	36.7	43.3	27.5	32.5	38.2	44.6	50.0	39.1
QA-FewDet [18]	ICCV 2021	41.0	33.2	35.3	47.5	52.0	23.5	29.4	37.9	35.9	37.1	33.2	29.4	37.6	39.8	41.4	36.9
FSCE [7]	CVPR 2021	32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0	41.2
DeFRCN [13]	ICCV 2021	40.2	53.6	58.2	63.6	**66.5**	29.5	**39.7**	43.4	**48.1**	**52.8**	35.0	38.3	**52.9**	57.7	**60.8**	49.4
Meta Faster R-CNN [19]	AAAI 2022	40.2	30.5	33.3	42.3	46.9	26.8	32.0	39.0	37.7	37.4	34.0	32.5	34.4	42.7	44.3	36.9
Ours	This work	**44.0**	**57.2**	**59.0**	**64.3**	**66.5**	**31.6**	39.2	**43.5**	47.1	51.1	**38.4**	**50.1**	52.1	**58.1**	59.8	**50.8**

overall average performance of our method is the best, with an improvement of 1.4%.

In the setting of 2-shot of split1, we visualiz the bounding boxes with confidence greater than 0.7, as shown in Figure 3. The successful and failure detection cases are shown to help us analyze the error types. In the successful cases, our model performs well in detection of some novel classes, particularly under challenging conditions such as complex background and low illumination. In the case of failure, there are more missed and false detections for the novel class of small objects and similar objects. The probably reasons are that the multi-scale feature extractor is not used in our approach and there are not many examples to compare in 2-shot.

TABLE II

Validation of FS-IPCL on individual modules.

Ablation choice/Metrics				1	2	3	5	10
ICL	PCL	cos	PCB	nAP50				
			✓	40.2	53.1	57.8	61.8	64.7
✓			✓	43.0	53.3	58.7	63.1	65.8
	✓		✓	42.3	55.5	57.9	63.0	65.6
✓		✓	✓	42.8	55.0	58.6	63.9	66.4
	✓	✓	✓	43.4	54.8	58.3	63.9	66.1
✓	✓	✓	✓	**44.0**	**57.2**	**59.0**	**64.3**	**66.5**

2) Ablation Study: **Ablation for each module of FS-IPCL**. The effectiveness of modules showing in the Table II, we confirm the effectiveness of the ICL and PCL modules separately. From the experimental results, we can see that both ICL and PCL improve the effect of using the model alone. However, from the last three rows of Table II, it can be seen that the combination of ICL and PCL has the best effect. It is probably because the ICL compares objects based on local instance structure, while PCL uses global semantics to construct of the object and compare global semantic structures information of the object. These two losses are complementary and are able to combine local contrast and global contrast.

Ablation for decoupling positive and negative sample strategy. To further verify the effectiveness of the decoupled positive and negative sample strategy, we conduct explicit experiments to show its performance. Better performance is achieved by decoupling the positive and negative samples in the contrastive loss, as demonstrated in Table III.

TABLE III

Ablation experiments for decoupling strategy for positive and negative samples in FS-IPCL.

Decouple the positive and negative samples	Novel Spilt 1		
	3	5	10
×	58.3	62.6	65.2
✓	**58.6**	**63.9**	**66.4**

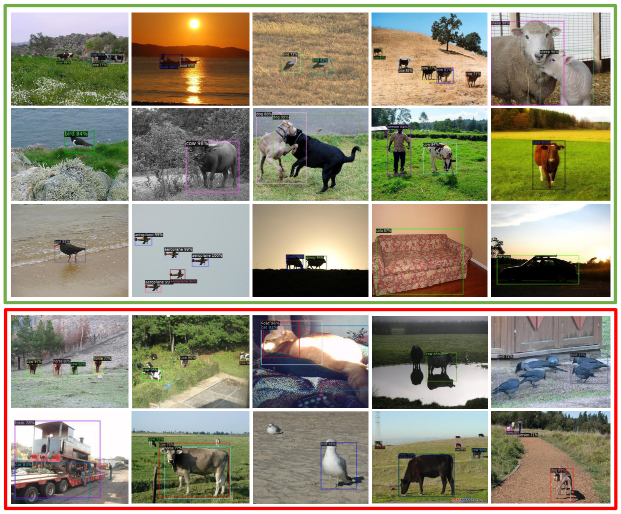

Fig. 3. We present the visualization results of the 2-shot object detection on the Pascal VOC dataset, where the bounding box score is greater than 0.7. The green box indicates successful cases of our FS-IPCL, while the red box represents the failure cases.

Ablation for prototype updates. We validate the effectiveness of our prototype updates. Meanwhile, we adopt IoU threshold to control the RoI features that need to be propagated for information. We test the effect of prototype update for RoI features with threshold greater than 0.7 and threshold greater than 0.5, and the experimental results are shown in Table IV. It turns out that the dynamic update of the prototype using RoI features with high IoU is better. This may be due to the fact that features with low IoU may contain a lot of background noise, causing the updated prototype feature embeddings to deviate from its class semantic center.

TABLE IV
ABLATION EXPERIMENTS FOR PROTOTYPE UPDATES IN FS-IPCL.

Prototype updates	Threshold	Novel Spilt 1		
		3	5	10
×	-	56.4	64.2	65.5
✓	$u_i > 0.5$	57.5	63.9	66.0
✓	$u_i > 0.7$	**59.0**	**64.3**	**66.5**

C. Experiments on MS-COCO

With more categories than Pascal VOC, the MS-COCO dataset contains more complex scenarios, making the detection of the model on the MS-COCO dataset more challenging. We validate the effectiveness of our model on 10 and 30 shots of MS-COCO. The results of the model's detection in the novel class are presented in Table V. It can be seen that our method has an accuracy improvement of 0.6%∼3.1% compared with most of the current methods.

TABLE V
RESULTS FOR 10 AND 30 SHOT ON THE MS-COCO DATASET.

Method	10-shot		30-shot	
	nAP	nAP75	nAP	nAP75
FRCN-ft [16]	5.5	5.5	7.4	7.4
FSRW [12]	5.6	4.6	9.1	7.6
TFA w/cos [4]	9.1	8.8	12.1	12.0
Viewpoint [17]	10.7	6.5	15.9	15.1
MPSR [5]	9.8	9.7	14.1	14.2
QA-FewDet [18]	10.2	9.0	11.5	10.3
FSCE [7]	11.1	9.8	15.3	14.2
Meta Faster R-CNN [19]	9.7	9.0	11.3	10.6
ours	**12.7**	**12.9**	**16.5**	**16.8**

V. CONCLUSION

In this work, we have proposed a novel approach by combining instance-wise contrastive learning and prototype contrastive learning to learn a balanced feature space. The instance-wise and prototype contrastive loss have been utilized to maximize the inter-class distance and minimize the intra-class distance of the novel class, and the discriminative features of the novel class have been obtained to improve the classification performance of the model for the novel class. In order to better utilize positive and negative samples for contrastive learning, a contrastive sample selecting scheme and a decoupling approach of positive and negative samples have been employed to improve the contrastive loss. Additionally, the prototypes have been dynamically updated using GCN to produce more representative prototype embeddings. Comparative experiments with several state-of-the-art methods based on meta-learning and fine-tuning have proved the proposed model always achieving competitive results. Our work takes a supervised contrastive learning approach to advance research in FSOD. In the future, we will further explore how to introduce unsupervised contrastive learning into FSOD to drive the development of this field.

ACKNOWLEDGMENT

The research reported in this article was supported by the Development Project of Ship Situational Intelligent Awareness System under the Grant MC-201920X01, the National Natural Science Foundation of China under the Grant No.61991410.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] W. Yang, S. Sheng, X. Luo, and S. Xie, "Geometric relation based point clouds classification and segmentation," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 11, p. e6845, 2022.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

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

[5] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 456–472.

[6] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6918–6928.

[7] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7352–7362.

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[9] J. Li, P. Zhou, C. Xiong, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[10] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*. Springer, 2022, pp. 668–684.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[12] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8420–8429.

[13] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8681–8690.

[14] J. Li, C. Xiong, and S. C. Hoi, "Mopro: Webly supervised learning with momentum prototypes," *arXiv preprint arXiv:2009.07995*, 2020.

[15] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[16] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9577–9586.

[17] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[18] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3263–3272.

[19] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 780–789.