Multi-Scale Positive Sample Refinement for Few-Shot Object Detection

Jiaxi Wu[1,2,3], Songtao Liu[1,2,3], Di Huang[1,2,3⋆], and Yunhong Wang[1,3]

[1] BAIC for BDBC, Beihang University, Beijing 100191, China
[2] SKLSDE, Beihang University, Beijing 100191, China
[3] SCSE, Beihang University, Beijing 100191, China
{wujiaxi,liusongtao,dhuang,yhwang}@buaa.edu.cn

Abstract. Few-shot object detection (FSOD) helps detectors adapt to unseen classes with few training instances, and is useful when manual annotation is time-consuming or data acquisition is limited. Unlike previous attempts that exploit few-shot classification techniques to facilitate FSOD, this work highlights the necessity of handling the problem of scale variations, which is challenging due to the unique sample distribution. To this end, we propose a Multi-scale Positive Sample Refinement (MPSR) approach to enrich object scales in FSOD. It generates multi-scale positive samples as object pyramids and refines the prediction at various scales. We demonstrate its advantage by integrating it as an auxiliary branch to the popular architecture of Faster R-CNN with FPN, delivering a strong FSOD solution. Several experiments are conducted on PASCAL VOC and MS COCO, and the proposed approach achieves state of the art results and significantly outperforms other counterparts, which shows its effectiveness. Code is available at https://github.com/jiaxi-wu/MPSR.

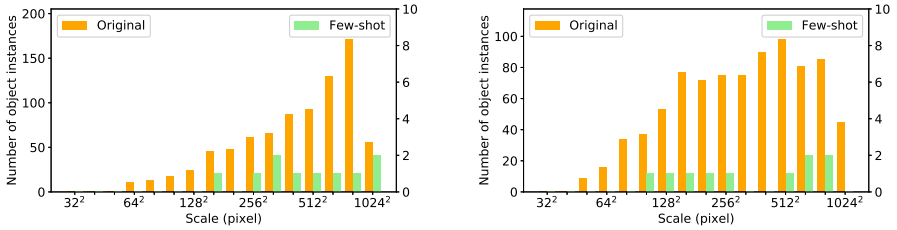
Keywords: Few-Shot Object Detection, Multi-Scale Refinement

1 Introduction

Object detection makes great progress these years following the success of deep convolutional neural networks (CNN) [32,15,11,31,3]. These CNN based detectors generally require large amounts of annotated data to learn extensive numbers of parameters, and their performance significantly drops when training data are inadequate. Unfortunately, for object detection, labeling data is quite expensive and the samples of some object categories are even hard to collect, such as endangered animals or tumor lesions. This triggers considerable attentions to effective detectors dealing with limited training samples. Few-shot learning is a popular and promising direction to address this issue. However, the overwhelming majority of the existing few-shot investigations focus on object/image classification, while the efforts on the more challenging *few-shot object detection* (FSOD) task are relatively rare.

⋆ indicates corresponding author (ORCID: 0000-0002-2412-9330).

With the massive parameters of CNN models, training detectors from scratch with scarce annotations generally incurs a high risk of overfitting. Preliminary research [4] tackles this problem in a transfer learning paradigm. Given a set of base classes with sufficient annotations and some novel classes with only a few samples, the goal is to acquire meta-level knowledge from base classes and then apply it to facilitating few-shot learning in detection of novel classes. Subsequent works [17,8,16,41] strengthen this pipeline by bringing more advanced methods on few-shot image classification, and commonly emphasize to improve classification performance of Region-of-Interest (RoI) in FSOD by using metric learning techniques. With elaborately learned representations, they ameliorate the similarity measurement between RoIs and marginally annotated instances, reporting better detection results. Meanwhile, [16,41] also attempt to deliver more general detectors, which account for all the classes rather than the novel ones only, by jointly using their samples in the training phase.



(a) Bus (b) Cow

Fig. 1. Illustration of scale distributions of two specific classes: (a) bus and (b) cow, in PASCAL VOC (Original) and a 10-shot subset (Few-shot). Images are resized with the shorter size at 800 pixels for statistics

The prior studies demonstrate that the FSOD problem can be alleviated in a similar manner as few-shot image classification. Nevertheless, object detection is much more difficult than image classification, as it involves not only classification but also localization, where the threat of varying scales of objects is particularly evident. The scale invariance has been widely explored in generic supervised detectors [11,33,34,18], while it remains largely intact in FSOD. Moreover, restricted by the quantity of annotations, this scale issue is even more tricky. As shown in Fig. 1, the lack of labels of novel classes leads to a sparse scale space (green bars) which may be totally divergent from the original distribution (yellow bars) of abundant training data. One could assume to make use of current effective solutions from generic object detection to enrich the scale space. For instance, Feature Pyramid Network (FPN), which builds multi-scale feature maps to detect objects at different scales, applies to situations where significant scale variations exist [22]. This universal property does contribute to FSOD, but it will not mitigate the difference of the scale distribution in the data of novel classes.

Regarding image pyramids [14,11], they build multi-scale representations of an image and allow detectors to capture objects in it at different scales. Although they are expected to narrow such a gap between the two scale distributions, the case is not so straightforward. Specifically, multi-scale inputs result in an increase in improper negative samples due to anchor matching. These improper negative samples contain a part of features belonging to the positive samples, which interferes their recognition. With abundant data, the network learns to extract diverse contexts and suppress the improper local patterns. But it is harmful to FSOD where both semantic and scale distributions are sparse and biased.

In this work, we propose a Multi-scale Positive Sample Refinement (MPSR) approach to few-shot object detection, aiming at solving its unique challenge of sparse scale distribution. We take the reputed Faster R-CNN as the basic detection model and employ FPN in the backbone network to improve its tolerance to scale variations. We then exploit an auxiliary refinement branch to generate multi-scale positive samples as object pyramids and further refine the prediction. This additional branch shares the same weights with the original Faster R-CNN. During training, this branch classifies the extracted object pyramids in both the Region Proposal Network (RPN) and the detector head. To keep scale-consistent prediction without introducing more improper negatives, we abandon the anchor matching rules and adaptively assign the FPN stage and spatial locations to the object pyramids as positives. It is worth noting that as we use no extra weights in training, our method achieves remarkable performance gains in an inference cost-free manner and can be conveniently deployed on different detectors.

The contributions of this study are three-fold:

1. To the best of our knowledge, it is the first work to discuss the scale problem in FSOD. We reveal the sparsity of scale distributions in FSOD with both quantitative and qualitative analysis.
2. To address this problem, we propose the MPSR approach to enrich the scale space without largely increasing improper negatives.
3. Comprehensive experiments are carried out, and significant improvements from MPSR demonstrate its advantage.

2 Related Work

Few-Shot Image Classification. There are relatively many historical studies in the area of few-shot image classification that targets recognition of objects with only a handful of images in each class [20,27]. [9] learns to initialize weights that effectively adapt to unseen categories. [1,28] aim to predict network parameters without heavily training on novel images. [19,38,35] employ metric learning to replace linear classifiers with learnable metrics for comparison between query and support samples. Although few-shot image classification techniques are usually used to advance the phase of RoI classification in FSOD, they are different tasks, as FSOD has to consider localization in addition.

Generic Object Detection. Recent object detection architectures are mainly divided into two categories: one-stage detectors and two-stage detectors. One-stage detectors use a single CNN to directly predict bounding boxes [29,30,25,24], and two-stage ones first generate region proposals and then classify them for decision making [12,11,31]. Apart from network design, scale invariance is an important aspect to detectors and many solutions have recently been proposed to handle scale changes [22,33,34,18]. For example, [22] builds multi-scale feature maps to match objects at different scales. [33] performs scale normalization to detect scale-specific objects and adopts image pyramids for multi-scale detection. These studies generally adapt to alleviate large size differences of objects. Few-shot object detection suffers from scale variations in a more serious way where a few samples sparsely distribute in the scale space.

Object Detection with Limited Annotations. To relieve heavy annotation dependence in object detection, there exist two main directions without using external data. One is weakly-supervised object detection, where only image-level labels are provided and spatial supervision is unknown [2]. Research basically concentrates on how to rank and classify region proposals with only coarse labels through multiple instance learning [36,37,39]. Another is semi-supervised object detection that assumes abundant images are available while the number of bounding box annotations is limited [26]. In this case, previous studies confirm the effectiveness of adopting extra images by pseudo label mining [5,10] or multiple instance learning [21]. Both the directions reduce manual annotation demanding to some extent, but they heavily depend on the amount of training images. They have the difficulty in dealing with constrained conditions where data acquisition is inadequate, *i.e.*, few-shot object detection.

Few-Shot Object Detection. Preliminary work [4] on FSOD introduces a general transfer learning framework and presents the Low-Shot Transfer Detector (LSTD), which reduces overfitting by adapting pre-trained detectors to few-shot scenarios with limited training images. Following this framework, RepMet [17] incorporates a distance metric learning classifier into the RoI classification head in the detector. Instead of categorizing objects with fully-connected layers, RepMet extracts representative embedding vectors by clustering and calculates distances between query and annotated instances. [8] is motivated by [19] which scores the similarity in a siamese network and computes pair-wise object relationship in both the RPN and the detection head. [16] is a single-stage detector combined with a meta-model that re-weights the importance of features from the base model. The meta-model encodes class-specific features from annotated images at a proper scale, and the features are viewed as reweighting coefficients and fed to the base model. Similarly, [41] delivers a two-stage detection architecture and re-weights RoI features in the detection head. Unlike previous studies where spatial influence is not considered, we argue that scale invariance is a challenging issue to FSOD, as the samples are few and their scale distribution is sparse. We improve the detector by refining object crops rather than masked

images [16,41] or siamese inputs [8] for additional training, which enriches the scale space and ensures the detector being fully trained at all scales.

3 Background

Before introducing MPSR, we briefly review the standard protocols and the basic detector we adopt for completeness. As it is the first work that addresses the challenge of sparse scale distribution in FSOD, we conduct some preliminary attempts with the current effective methods from generic object detection (*i.e.*, FPN and image pyramids) to enrich the scale space and discuss their limitations.

3.1 Baseline Few-Shot Object Detection

Few-Shot Object Detection Protocols. Following the settings in [16,41], object classes are divided into base classes with abundant data and novel classes with only a few training samples. The training process of FSOD generally adopts a two-step paradigm. During base training, the detection network is trained with a large-scale dataset that only contains base classes. Then the detection network is fine-tuned on the few-shot dataset, which only contains a very small number of balanced training samples for both base and novel classes. This two-step training schedule avoids the risk of overfitting with insufficient training samples on novel classes. It also prevents the detector from extremely imbalanced training if all annotations from both base and novel classes are exploited together [41]. To build the balanced few-shot dataset, [16] employs the k-shot sampling strategy, where each object class only has k annotated bounding boxes. Another work [4] collects k images for each class in the few-shot dataset. As k images actually contain an arbitrary number of instances, training and evaluation under this protocol tend to be unstable. We thus use the former strategy following [16].

Basic Detection Model. With the fast development in generic object detection, the base detector in FSOD has many choices. [16] is based on YOLOv2 [30], which is a single-stage detector. [41] is based on a classical two-stage detector, Faster R-CNN [31], and demonstrates that Faster R-CNN provides consistently better results. Therefore, we take the latter as our basic detection model. Faster R-CNN consists of the RPN and the detection head. For a given image, the RPN head generates proposals with objectness scores and bounding-box regression offsets. The RPN loss function is:

$$L_{RPN} = \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} L^i_{Bcls} + \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} L^i_{Preg}. \tag{1}$$

For the ith anchor in a mini-batch, L^i_{Bcls} is the binary cross-entropy loss over background and foreground and L^i_{Preg} is the smooth L_1 loss defined in [31]. N_{obj} is the total number of chosen anchors. These proposals are used to extract

RoI features and then fed to the detection (RoI) head that outputs class-specific scores and bounding-box regression offsets. The loss function is defined as:

$$L_{RoI} = \frac{1}{N_{RoI}} \sum_{i=1}^{N_{RoI}} L^i_{Kcls} + \frac{1}{N_{RoI}} \sum_{i=1}^{N_{RoI}} L^i_{Rreg}, \tag{2}$$

where L^i_{Kcls} is the log loss over K classes and N_{RoI} is the number of RoIs in a mini-batch. Different from the original implementation in [31], we employ a class-agnostic regression task in the detection head, which is the same as [4]. The total loss is the sum of L_{RPN} and L_{RoI}.

3.2 Preliminary Attempts

FPN for Multi-Scale Detection. As FPN is commonly adopted in generic object detection to address the scale variation issue [22,3], we first consider applying it to FSOD in our preliminary experiments. FPN generates several different semantic feature maps at different scales, enriching the scale space in features. Our experiments validate that it is still practically useful under the restricted conditions in FSOD. We thus exploit Faster R-CNN with FPN as our second baseline. However, FPN does not change the distribution in the data of novel classes and the sparsity of scale distribution remains unsolved in FSOD.

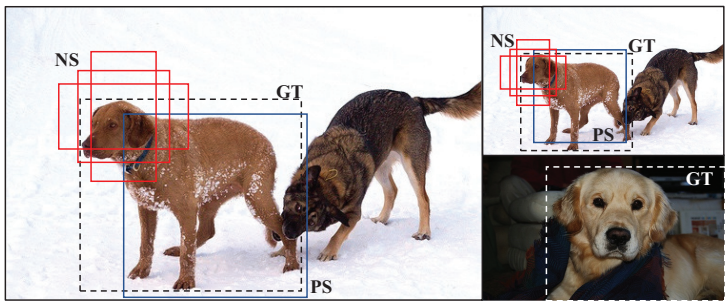


Fig. 2. An example of improper negative samples in FSOD. Negative samples (NS), positive samples (PS) and ground-truth (GT) bounding boxes are annotated. The improper negative samples significantly increase as more scales are involved (top right), while they may even be true positives in other contexts (bottom right)

Image Pyramids for Multi-Scale Training. To enrich object scales, we then consider a multi-scale training strategy which is also widely used in generic object detection for multi-scale feature extraction [14,11] or data augmentation [30]. In few-shot object detection, image pyramids enrich object scales as data augmentation and the sparse scale distribution can be theoretically solved. However, this

multi-scale training strategy acts differently in FSOD with the increasing number of improper negative samples. As in Fig. 2, red bounding boxes are negative samples in training while they actually contain part of objects and may even be true positive samples in other contexts (as in bottom right). These improper negative samples require sufficient contexts and clues to suppress, inhibiting being mistaken for potential objects. Such an interference is trivial when abundant annotations are available, but it is quite harmful to the sparse and biased distribution in FSOD. Moreover, with multi-scale training, a large number of extra improper negative samples are introduced, which further hurts the performance.

4 Multi-Scale Positive Sample Refinement

4.1 Multi-Scale Positive Sample Refinement Branch

Motivated by the above discussion, we employ FPN in the backbone of Faster R-CNN as the advanced version of baseline. To enrich scales of positive samples without largely increasing improper negative samples, we extract each object independently and resize them to various scales, denoted as object pyramids. Specifically, each object is cropped by a square window (whose side is equal to the longer side of the bounding box) with a minor random shift. It is then resized to $\{32^2, 64^2, 128^2, 256^2, 512^2, 800^2\}$ pixels, which is similar to anchor design.

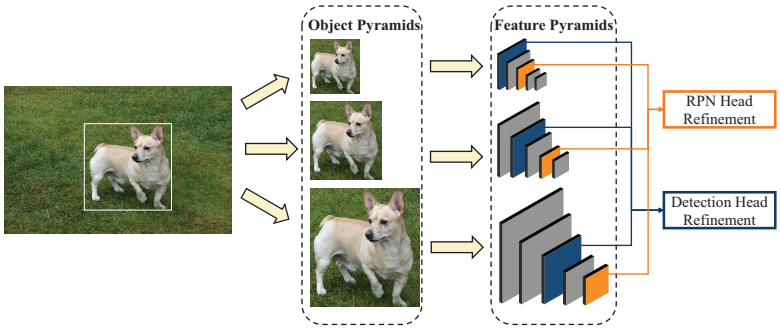


Fig. 3. Multi-scale positive sample feature extraction. The positive sample is extracted and resized to various scales. Specific feature maps from FPN are selected for refinement

In object pyramids, each image only contains a single instance, which is inconsistent to the standard detection pipeline. Therefore, we propose an extra positive sample refinement branch to adaptively project the object pyramids into the standard detection network. For a given object, the standard FPN pipeline samples the certain scale level and the spatial locations as positives for training, operated by anchor matching. However, performing anchor matching on cropped single objects is wasteful and also incurs more improper negatives that hurt the

performance for FSOD. As shown in Fig. 3, instead of anchor matching, we manually select the corresponding scale level of feature maps and the fixed center locations as positives for each object, keeping it consistent with the standard FPN assigning rules. After selecting specific features from these feature maps, we feed them directly to the RPN head and the detection head for refinement.

Table 1. FPN feature map selection for different object scales. For each object, two specific feature maps are activated, fed to RPN and detection (RoI) heads respectively

	32^2	64^2	128^2	256^2	512^2	800^2
RPN	P_2	P_3	P_4	P_5	P_6	P_6
RoI	P_2	P_2	P_2	P_3	P_4	P_5

In the RPN head, the multi-scale feature maps of FPN $\{P_2, P_3, P_4, P_5, P_6\}$ represent anchors whose areas are $\{32^2, 64^2, 128^2, 256^2,\ 512^2\}$ pixels respectively. For a given object, only one feature map with the consistent scale is activated, as shown in Table 1. To simulate that each proposal is predicted by its center location in RPN, we select centric 2^2 features for object refinement. We also put anchors with $\{1:2, 1:1, 2:1\}$ aspect ratios on the sampled locations. These selected anchors are viewed as positives for the RPN classifier.

To extract RoI features for the detection head, only $\{P_2, P_3, P_4, P_5\}$ are used and the original RoI area partitions in the standard FPN pipeline are: $\left(0^2, 112^2\right)$, $\left[112^2, 224^2\ \right)$, $\left[224^2, 448^2\ \right)$, $\left[448^2, \infty\ \right)$ [22]. We also select one feature map at a specific scale for each object to keep the scale consistency, as shown in Table 1. As the randomly cropped objects tend to have larger sizes than the orignal ground truth bounding boxes, we slightly increase the scale range of each FPN stage for better selection. Selected feature maps are adaptively pooled to the same RoI size and fed to the RoI classifier.

4.2 Framework

As shown in Fig. 4, the whole detection framework for training consists of Faster R-CNN with FPN and the refinement branch working in parallel while sharing the same weights. For a given image, it is processed by the backbone network, RPN, RoI Align layer, and the detection head in the standard two-stage detection pipeline [31]. Simultaneously, an independent object extracted from the original image is resized to different scales as object pyramids. The object pyramids are fed into the detection network as described above. The outputs from RPN and detection heads in the MPSR branch include objectness scores and class-specific scores similar to the definitions in Section 3.1. The loss function of the RPN head containing Faster R-CNN and the MPSR branch is defined as:

$$L_{RPN} = \frac{1}{N_{obj}+M_{obj}} \sum_{i=1}^{N_{obj}+M_{obj}} L^i_{Bcls} + \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} L^i_{Preg}, \tag{3}$$

where M_{obj} is the number of selected positive anchor samples for refinement. The loss function of the detection head is defined as:

$$L_{RoI} = \frac{1}{N_{RoI}} \sum_{i=1}^{N_{RoI}} L^i_{Kcls} + \frac{\lambda}{M_{RoI}} \sum_{i=1}^{M_{RoI}} L^i_{Kcls} + \frac{1}{N_{RoI}} \sum_{i=1}^{N_{RoI}} L^i_{Rreg}, \tag{4}$$

where M_{RoI} is the number of selected RoIs in MPSR. Unlike the RPN head loss where M_{obj} is close to N_{obj}, the number of positives from object pyramids is quite small compared to N_{RoI} in the RoI head. We thus add a weight parameter λ to the RoI classification loss of the positives from MPSR to adjust its magnitude, which is set to 0.1 by default. After the whole network is fully trained, the extra MPSR branch is removed and only Faster R-CNN with FPN is used for inference. Therefore, the MPSR approach that we propose benefits FSOD training without extra time cost at inference.

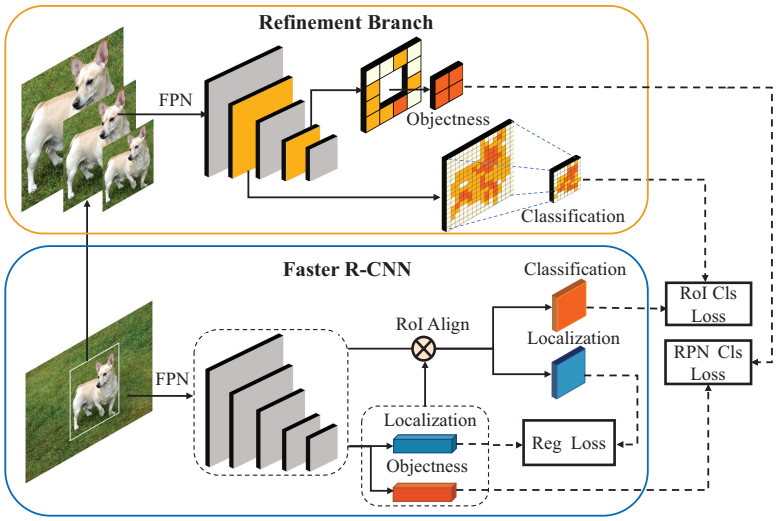


Fig. 4. MPSR architecture. On an input image to Faster R-CNN, the auxiliary branch extracts samples and resizes them to different scales. Each sample is fed to the FPN and specific features are selected to refine RPN and RoI heads in Faster R-CNN

5 Experiments

5.1 Datasets and Settings

We evaluate our method on the PASCAL VOC 2007 [7], 2012 [6] and MS COCO [23] benchmarks. For fair quantitative comparison with state of the art (SOTA) methods, we follow the setups in [16,41] to construct few-shot detection datasets.

PASCAL VOC. Our networks are trained on the modified VOC 2007 trainval and VOC 2012 trainval sets. The standard VOC 2007 test set is used for evaluation. The evaluation metric is the mean Average Precision (mAP). Both the trainval sets are split by object categories, where 5 are randomly chosen as novel classes and the left 15 are base classes. Here we follow [16] to use the same three class splits, where the unseen classes are {"bird", "bus", "cow", "motorbike" ("mbike"), "sofa"}, {"aeroplane" ("aero"), "bottle", "cow", "horse", "sofa"}, {"boat", "cat", "motorbike", "sheep", "sofa"}, respectively. For FSOD experiments, the few-shot dataset consists of images where only k object instances are available for each category and k is set as 1/3/5/10.

MS COCO. COCO has 80 object categories, where the 20 categories overlapped with PASCAL VOC are denoted as novel classes. 5,000 images from the val set, denoted as minival, are used for evaluation while the left images in the train and val sets are used for training. Base and few-shot dataset construction is the same as that in PASCAL VOC except that k is set as 10/30.

Implementation Details. We train and test detection networks on images of a single scale. We resize input images so that their shorter sides are set to 800 pixels and the longer sides are less than 1,333 pixels while maintaining the aspect ratio. Our backbone is ResNet-101 [15] with the RoI Align [13] layer and we use the weights pre-trained on ImageNet [32] in initialization. For efficient training, we randomly sample one object to generate the object pyramid for each image. After training on base classes, only the last fully-connected layer (for classification) of the detection head is replaced. The new classification layer is randomly initialized and none of the network layers is frozen during few-shot fine-tuning. We train our networks with a batchsize of 4 on 2 GPUs, 2 images per GPU. We run the SGD optimizer with the momentum of 0.9 and the parameter decay of 0.0001. For base training on VOC, models are trained for 240k, 8k, and 4k iterations with learning rates of 0.005, 0.0005 and 0.00005 respectively. For few-shot fine-tuning on VOC, we train models for 1,300, 400, 300 iterations and the learning rates are 0.005, 0.0005 and 0.00005, respectively. Models are trained on base COCO classes for 56k, 14k, and 10k iterations. For COCO few-shot fine-tuning, the 10-shot dataset requires 2,800, 700, and 500 iterations, while the 30-shot dataset requires 5,600, 1,400, 1,000 iterations.

5.2 Results

We compare our results with two baseline methods (denoted as Baseline and Baseline-FPN) as well as two SOTA few-shot detection counterparts. Baseline and Baseline-FPN are our implemented Faster R-CNN and Faster R-CNN with FPN described in Section 3. YOLO-FS [16] and Meta R-CNN [41] are the SOTA few-shot detectors based on DarkNet-19 and ResNet-101, respectively. It should be noted that due to better implementation and training strategy, our baseline achieves higher performance than SOTA, which is also confirmed by the very recent work [40].

Table 2. Comparison of different methods in terms of mAP (%) of novel classes using the three splits on the VOC 2007 test set

	Class Split 1				Class Split 2				Class Split 3			
Method/Shot	1	3	5	10	1	3	5	10	1	3	5	10
YOLO-FS [16]	14.8	26.7	33.9	47.2	15.7	22.7	30.1	39.2	19.2	25.7	40.6	41.3
Meta R-CNN [41]	19.9	35.0	45.7	51.5	10.4	29.6	34.8	45.4	14.3	27.5	41.2	48.1
Baseline	24.5	40.8	44.6	47.9	16.7	34.9	37.0	40.9	27.3	36.3	41.2	45.2
Baseline-FPN	25.5	41.1	49.6	56.9	15.5	37.7	38.9	43.8	29.9	37.9	46.3	47.8
MPSR (ours)	**41.7**	**51.4**	**55.2**	**61.8**	**24.4**	**39.2**	**39.9**	**47.8**	**35.6**	**42.3**	**48.0**	**49.7**

PASCAL VOC. MPSR achieves 82.1%/82.7%/82.9% on base classes of three splits respectively before few-shot fine-tuning. The main results of few-shot experiments on VOC are summarized in Table 2. It can be seen from this table that the results of the two baselines (*i.e.* Baseline and Baseline-FPN) are close to each other when the number of instances is extremely small (*e.g.* 1 or 3), and Baseline-FPN largely outperforms the other as the number of images increases. This demonstrates that FPN benefits few-shot object detection as in generic object detection. Moreover, our method further improves the performance of Baseline-FPN with any number of training samples in all the three class splits. Specifically, by solving the sparsity of object scales, we achieve a significant increase in mAP compared to the best scores of the two baselines, particularly when training samples are extremely scarce, *e.g.* 16.2% on 1-shot split-1. It clearly highlights the effectiveness of the extra MPSR branch. Regarding other counterparts [16,41], the proposed approach outperforms them by a large margin, reporting the state of the art scores on this dataset.

Table 3. AP (%) of each novel class on the 3-/10-shot VOC dataset of the first class split. mAP (%) of novel classes and base classes are also presented

		Novel Classes					Mean	
Shot	Method	bird	bus	cow	mbike	sofa	Novel	Base
3	YOLO-FS [16]	26.1	19.1	40.7	20.4	27.1	26.7	64.8
	Meta R-CNN[41]	30.1	44.6	50.8	38.8	10.7	35.0	64.8
	Baseline	34.9	26.9	53.3	50.8	38.2	40.8	45.2
	Baseline-FPN	32.6	29.4	45.5	56.2	41.7	41.1	66.2
	MPSR (ours)	**35.1**	**60.6**	**56.6**	**61.5**	**43.4**	**51.4**	**67.8**
10	YOLO-FS [16]	30.0	62.7	43.2	60.6	39.6	47.2	63.6
	Meta R-CNN [41]	**52.5**	55.9	52.7	54.6	41.6	51.5	67.9
	Baseline	38.6	48.6	51.6	57.2	43.4	47.9	47.8
	Baseline-FPN	41.8	68.4	61.7	66.8	45.8	56.9	70.0
	MPSR (ours)	48.3	**73.7**	**68.2**	**70.8**	**48.2**	**61.8**	**71.8**

Following [16,41], we display the detailed results of 3-/10-shot detection in the first split on VOC in Table 3. Consistently, our Baseline-FPN outperforms the existing methods on both the novel and base classes. This confirms that FPN addresses the scale problem in FSOD to some extent. Furthermore, our method improves the accuracies of Baseline-FPN in all the settings by integrating MPSR, illustrating its advantage.

MS COCO. We evaluate the method using 10-/30-shot setups on MS COCO with the standard COCO metrics. The results on novel classes are provided in Table 4. Although COCO is quite challenging, we still achieve an increase of 0.4% on 30-shot compared with Baseline-FPN while boosting the SOTA mAP from 12.4% (Meta R-CNN) to 14.1%. Specifically, our method improves the recognition of small, medium and large objects simultaneously. This demonstrates that our balanced scales of input objects are effective.

Table 4. AP (%) and AR (%) of 10-/30-shot scores of novel classes on COCO minival

Shot	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
10	YOLO-FS [16]	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
	Meta R-CNN [41]	8.7	**19.1**	6.6	2.3	7.7	14.0	12.6	17.8	17.9	**7.8**	15.6	27.2
	Baseline	8.8	18.7	7.1	2.9	8.1	15.0	12.9	17.2	17.2	4.1	14.2	29.1
	Baseline-FPN	9.5	17.3	9.4	2.7	8.4	15.9	14.8	20.6	20.6	4.7	19.3	33.1
	MPSR (ours)	**9.8**	17.9	**9.7**	**3.3**	**9.2**	**16.1**	**15.7**	**21.2**	**21.2**	4.6	**19.6**	**34.3**
30	YOLO-FS [16]	9.1	19.0	7.6	0.8	4.9	16.8	13.2	17.7	17.8	1.5	10.4	33.5
	Meta R-CNN [41]	12.4	25.3	10.8	2.8	11.6	19.0	15.0	21.4	21.7	**8.6**	20.0	32.1
	Baseline	12.6	**25.7**	11.0	3.2	11.8	20.7	15.9	21.8	21.8	5.1	18.0	36.9
	Baseline-FPN	13.7	25.1	13.3	3.6	12.5	**23.3**	**17.8**	**24.7**	**24.7**	5.4	**21.6**	**40.5**
	MPSR (ours)	**14.1**	25.4	**14.2**	**4.0**	**12.9**	23.0	17.7	24.2	24.3	5.5	21.0	39.3

MS COCO to PASCAL VOC. We conduct cross-dataset experiments on the standard VOC 2007 test set. In this setup, all the models are trained on the base COCO dataset and finetured with 10-shot objects in novel classes on VOC. Results of Baseline and Baseline-FPN are 38.5% and 39.3% respectively. They are worse than 10-shot results only trained on PASCAL VOC due to the large domain shift. Cross-dataset results of YOLO-FS and Meta R-CNN are 32.3% and 37.4% respectively. Our MPSR achieves 42.3%, which indicates that our method has better generalization ability in cross-domain situations.

5.3 Analysis of Sparse Scales

We visualize the scale distribution of two categories on the original dataset (Pascal VOC) and 10-shot subset in Fig. 1. It is obvious that the scale distribution in the few-shot dataset is extremely sparse and distinct from the original ones.

Table 5. AP (%) on bus/cow class. Two 10-shot datasets are constructed on VOC split-1, where scales of instances are random or limited. Std over 5 runs are presented

	Bus		Cow	
Method	Random	Limited	Random	Limited
Baseline-FPN	68.4±0.6	39.5±1.3	61.7±0.9	39.9±1.2
MPSR (ours)	73.7±1.6	54.0±1.4	68.2±1.0	52.5±1.6

To quantitatively analyze the negative effect of scale sparsity, we evaluate detectors on two specific 10-shot datasets. We carefully select the bus and cow instances with the scale between 128^2 and 256^2 pixels to construct the "limited" few-shot datasets. As shown in Table 5, such extremely sparse scales lead to a significant drop in performance (*e.g.* for bus, -28.9% on Baseline-FPN). Therefore, it is essential to solve the extremely sparse and biased scale distribution in FSOD. With our MPSR, the reduction of performance is relieved.

Table 6. mAP (%) comparison of novel/base classes on VOC split-1: Baseline-FPN, SNIPER [34], Baseline-FPN with scale augmentation/image pyramids and MPSR

	Novel			Base		
Method/Shot	1	3	5	1	3	5
Baseline-FPN	25.5	41.1	49.6	56.9	66.2	67.9
SNIPER [34]	1.4	21.0	39.7	**67.8**	**74.8**	**76.2**
Scale Augmentation	29.8	44.7	49.8	52.7	67.1	68.8
Image Pyramids	29.5	48.4	50.4	58.1	67.5	68.3
MPSR (ours)	**41.7**	**51.4**	**55.2**	59.4	67.8	68.4

As in Table 6, we compare MPSR with several methods that are used for scale invariance. SNIPER [34] shows a lower accuracy on novel classes and a higher accuracy on base classes than the baseline. As SNIPER strictly limits the scale range in training, it actually magnifies the sparsity of scales in FSOD. Such low performance also indicates the importance of enriching scales. We also evaluate the scale augmentation and image pyramids with a shorter side of $\{480, 576, 688, 864, 1200\}$ [14]. We can see that our MPSR achieves better results than those two multi-scale training methods on the novel classes. When only one instance is available for each object category, our method exceeds multi-scale training by $\sim$12%, demonstrating its superiority.

5.4 Ablation Studies

We conduct some ablation studies to verify the effectiveness of the proposed manual selection and refinement method in Table 7.

Table 7. mAP (%) of MPSR with different settings of novel classes on VOC split-1

Baseline FPN	Object Pyramids	Manual Selection	Refinement RPN	Refinement RoI	Shot 1	Shot 3	Shot 5
✓					25.5	41.1	49.6
✓	✓		✓	✓	30.8	43.6	49.6
✓	✓	✓	✓		36.7	48.0	54.4
✓	✓	✓		✓	33.7	48.2	54.7
✓	✓	✓	✓	✓	41.7	51.4	55.2

Manual Selection. From the first two lines in Table 7, we see that applying anchor matching to object pyramids on both RPN and RoI heads achieves better performance than Baseline-FPN. However, when compared to the last three lines with manual selection rules, anchor matching indeed limits the benefits of object pyramids, as it brings more improper negative samples to interfere few-shot training. It confirms the necessity of the proposed manual refinement rules.

RPN and Detection Refinement. As in the last three lines of Table 7, we individually evaluate RPN refinement and detection (RoI) refinement to analyze their credits in the entire approach. Models with only the RPN and RoI refinement branches exceed Baseline-FPN in all the settings, which proves their effectiveness. Our method combines them and reaches the top score, which indicates that the two branches play complementary roles.

6 Conclusions

This paper targets the scale problem caused by the unique sample distribution in few-shot object detection. To deal with this issue, we propose a novel approach, namely multi-scale positive sample refinement. It generates multi-scale positive samples as object pyramids and refines the detectors at different scales, thus enlarging the scales of positive samples while limiting improper negative samples. We further deliver a strong FSOD solution by integrating MPSR to Faster R-CNN with FPN as an auxiliary branch. Experiments are extensively carried out on PASCAL VOC and MS COCO, and the proposed approach reports better scores compared to current state of the arts, which shows its advantage.

Acknowledgment

This work is funded by the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2019ZX-03) and the Fundamental Research Funds for the Central Universities.

References

1. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P.H.S., Vedaldi, A.: Learning feed-forward one-shot learners. In: Advances in Neural Information Processing Systems (NIPS) (2016)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
4. Chen, H., Wang, Y., Wang, G., Qiao, Y.: LSTD: A low-shot transfer detector for object detection. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. Dong, X., Zheng, L., Ma, F., Yang, Y., Meng, D.: Few-example object detection with model communication. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
6. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision (2015)
7. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. International Journal of Computer Vision (2010)
8. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning (ICML) (2017)
10. Gao, J., Wang, J., Dai, S., Li, L.J., Nevatia, R.: Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In: IEEE International Conference on Computer Vision (ICCV) (2019)
11. Girshick, R.B.: Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV) (2015)
12. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV) (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: IEEE International Conference on Computer Vision (ICCV) (2019)
17. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

18. Kim, Y., Kang, B., Kim, D.: SAN: learning relationship between convolutional features for multi-scale object detection. In: European Conference on Computer Vision (ECCV) (2018)
19. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML DeepLearning workshop (2015)
20. Li, F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2006)
21. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L., Fei-Fei, L.: Thoracic disease identification and localization with limited supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
24. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: European Conference on Computer Vision (ECCV) (2018)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: European Conference on Computer Vision (ECCV) (2016)
26. Misra, I., Shrivastava, A., Hebert, M.: Watch and learn: Semi-supervised learning of object detectors from videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
27. Munkhdalai, T., Yu, H.: Meta networks. In: International Conference on Machine Learning (ICML) (2017)
28. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
29. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
30. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
31. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) (2015)
33. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection-SNIP. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
34. Singh, B., Najibi, M., Davis, L.S.: SNIPER: efficient multi-scale training. In: Advances in Neural Information Processing Systems (NIPS) (2018)
35. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
36. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

37. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.L.: Weakly supervised region proposal network and object detection. In: European Conference on Computer Vision (ECCV) (2018)
38. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems (NIPS) (2016)
39. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-MIL: continuation multiple instance learning for weakly supervised object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
40. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. In: International Conference on Machine Learning (ICML) (2020)
41. 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: IEEE International Conference on Computer Vision (ICCV) (2019)