Open-World Few-Shot Object Detection

Abstract. General object detection has made significant progress under the close-set setting. However, the detector can only support a fixed set of categories and fail to identify unknown objects in real-world scenarios. Therefore, class-agnostic object detection (CAOD) has recently attracted much attention, aiming to localize both known and unknown objects in the image. Since CAOD utilizes the binary label to train the detector, lacking multi-class classification information, and is also incapable to further generalize quickly to the unknown objects of interest, the scalability of this task is limited in more downstream applications. In this paper, we propose a new task termed **O**pen-World **F**ew-Shot **O**bject **D**etection (OFOD), extending class-agnostic object detection with the few-shot learning ability. Compared with CAOD, OFOD can accurately detect unknown objects with only a few examples. Besides, we propose a new model termed OFDet, built upon a class-agnostic object detector under the two-stage fine-tuning paradigm. OFDet consists of three key components, Class-agnostic Localization Module (CALM) that generates class-agnostic proposals, Base Classification Module (BCM) that classifies objects from classes features, and Novel Detection Module (NDM) that learns to detect novel objects. OFDet detects the novel classes in NDM and localizes the potential unknown proposals in CALM. Furthermore, an Unknown Proposals Selection algorithm is proposed to select more accurate unknown objects. Extensive experiments are conducted on PASCAL VOC and COCO datasets under multiple tasks, CAOD, few-shot object detection (FSOD) and OFOD. The results show that OFDet performs well on the traditional FSOD and CAOD settings as well as the proposed OFOD setting. Specifically for OFOD, OFDet achieves state-of-the-art results on the average recall of unknown classes (32.5%) and obtains high average precision of novel classes (15.7%) under the 30-shot setting of COCO's unknown set 1.

Keywords: General Object Detection, Class-Agnostic Object Detection, Few-shot Learning, Unknown Proposal Selection.

1 Introduction

Deep neural networks have witnessed significant progress in object detection [1, 2, 3], which aims to localize and classify objects of interest in the image. However, the success of most modern object detectors is built on a close-set setting where the categories in the test set entirely depend on the ones used in the training process. Therefore, in more realistic scenarios, these detectors are unable to recognize the unseen categories in training.

In contrast, humans can recognize unseen objects similar to previous classes in new environments regardless of their special categories, leading us to research class-agnostic object detection (CAOD) [4, 5, 6]. As a sub-problem of open-set learning [7, 8], CAOD aims to localize all instances of objects in the image without learning to classify them. There are two scenarios for this task as follows. One performs as a significant pre-processing module in object detection, e.g., Region Proposal Network [2], to classify object proposals into foreground or background and provide more accurate proposals for subsequent modules. The other is to localize unseen objects in practical applications such as autonomous driving and robot navigation. Recently, Kim et al. proposed OLN [6] to learn better objectness cues (i.e., intersection of union and centerness [9]) for object localization, which generalizes well to unknown classes compared to previous works [2, 9].

Table 1. The effectiveness of Base Detection Module on 10-shot split 1 of PASCAL VOC.

Method	bAP_{50}	nAP_{50}
OFDet w/o BCM	35.5	30.2
OFDet w/ BCM	76.7	59.2

However, the scalability of such networks is still limited. Suppose we can employ the strong object locator as a pre-processing module, fine-tuned with only a few annotations of unknown objects of interest. In that case, the network will reduce the computational cost and be more flexible during the training process. To this end, we first attempt combining CAOD with few-shot object detection (FSOD) [10, 11, 12], which has attracted much attention recently in data-scarce scenarios. A few-shot detector can quickly generalize to novel classes with a few annotated examples by learning from abundant base examples. Therefore, a naive idea is to train a class-agnostic detector and then directly extend it with an R-CNN [13] module to classify known classes and detect unknown classes of interest (i.e., novel classes) with only a few annotations. We expect a competitive result compared to the model trained on abundant data. However, as shown in the first row of **Table 1**, the results of base and novel classes are neither satisfying. The reason may be that the class-agnostic detector is supervised with binary labels, lacking multi-class classification information. The newly added R-CNN module is under-fitted to base classes and further harms the generalization to novel classes. Therefore, we propose a simple yet effective module named Base Classification Module (BCM) to maintain the classification information of base classes. As shown in **Fig. 2** and **Table 1.** The effectiveness of Base Detection Module on 10-shot split 1 of PASCAL VOC., BCM boosts the average precision (AP) of base and novel classes by 30% approximately. In addition, since we still need to recognize the potential unknown objects in the second stage for further design, existing problem settings no longer meet our needs.

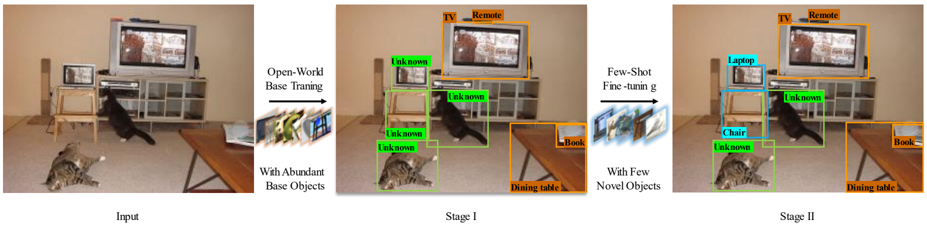

Fig. 1. An illustration of the proposed Open-world Few-shot Object Detection task. Given an input image, the model needs to detect a set of base classes (e.g., book, TV and remote) while localizing the unknown objects in stage I. In stage II, the model needs to recognize a set of novel classes (e.g., laptop and chair) from unknown objects with only a few annotated examples. Orange, green and blue denote base classes, unknown classes and novel classes, respectively.

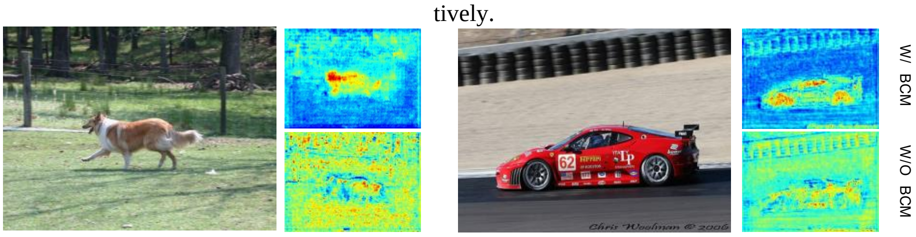

Fig. 2. The visualization of Base Classification Module

Here we propose a novel setting called Open-World Few-shot Object Detection (OFOD), shown in **Fig. 1**. OFOD not only needs to localize unseen objects by fully leveraging the representation learned from limited base classes in the first stage, but also detect novel categories with few examples in the next stage. We also propose a detector named OFDet for the OFOD task, which builds upon a class-agnostic object detector [6] under a two-stage fine-tuning paradigm [12]. OFDet consists of three novel modules: Class-Agnostic Localization Module (CALM), Base Classification Module (BCM) and Novel Detection Module (NDM). In the first stage, the class-agnostic proposals are utilized for training CALM consisting of location RoI head (L-RoI head) and the following box regression and localization quality layers. The classification features of base classes are extracted by the BCM, composed of Classification RoI head (C-RoI head) and the box classifier. In the second stage, the class-agnostic proposals are produced by CALM and pooled into a fixed-size feature map to perform box regression and novel classification in NDM.

We test our method on three tasks (CAOD, FSOD and OFOD) across two datasets, PASCAL VOC [14] and MS COCO [15]. For the FSOD task, our approach outperforms the baseline of transfer-learning approach under most settings. For the CAOD task, our network essentially improves the average recall of novel classes when we add a few novel annotations. For our proposed OFOD task, our network achieves state-of-the-art performance for the localization of unknown classes while keeping the capability of detecting known classes. Our main contributions are summarized as follows:

- We introduce a novel problem setting, Open-World Few-Shot Object Detection, which attempts to localize the unknown classes and detect the novel classes as FSOD task.
- We present a network named OFDet, based on class-agnostic object detection with a two-stage fine-tuning paradigm to address the challenge OFOD task. In OFDet, we propose three novel modules and an unknown proposal selection algorithm for known and unknown classes.
- Extensive experiments and analyses illustrate that our approach performs well across three tasks, demonstrating the effectiveness of our approach.

2 Related Work

General Object Detection. General Object Detection can be mainly divided into two branches, i.e., one-stage object detectors [1, 3, 16, 17] and two-stage object detectors [2, 18, 19]. Compared with two stage detectors, one stage detectors achieve fast inference but get low accuracy. Beyond the conventional form, DETR [20] based on transformer and set prediction paradigm, removes handcraft NMS [21] post-processing to achieve end-to-end object detection. However, these detectors are under a close-set condition and struggle to generalize to novel categories with limited annotations in more realistic scenarios.

Few-Shot Object Detection. Few-Shot Object Detection can be mainly divided into meta-learning based approaches and transfer-learning based approaches. Similar to few-shot classification [22, 23], early works mainly investigate the meta-learning approaches. Compared to general object detectors, meta-learning based approaches [11, 24, 25] introduce a meta-learner that generates class-attentive vectors to reweigh the query features for novel categories. In contrast, TFA [12] proposes a two-stage fine-tuning approach based on transfer learning, which removes the design of the meta-learner and only fine-tunes the last layer of the base model for novel classes. Recently, several follow-up works [26, 27, 28] based on TFA have been proposed. In this work, we adopt a similar approach to the transfer-learning based approaches. Despite the progress in detecting novel classes, few-shot detectors still struggle to generalize to unknown classes. In comparison, our method can detect novel objects and localize unknown objects in a unified framework.

Class-Agnostic Object Detection. Several works [4, 29] in the early research explore the paradigm of class-agnostic object proposals, which utilize hand-crafted heuristics to capture generalized and informative characteristics of objects [5, 30]. Recently those classical heuristics algorithms have been replaced by learning-based approaches [2, 19, 31] for better performance. Region Proposal Network (RPN) [2] is a representative architecture for generating class-agnostic proposals that are then used for downstream classification and segmentation tasks. There are some works [32, 33] following [2] to improve the localization quality and speed up the inference. Our study is more closely related to generalizing the model to localize unknown classes. Recently [6] proposes a pure classification-free object localization network (OLN) by estimating the objectness of each region, which outperforms existing methods on unseen classes. In our work, we build on OLN architecture for better performance in localizing the known and unknown objects. Compared to those class-agnostic detectors, our proposed method can detect novel classes from unknown objects with a few samples and get a higher average recall to solve the practical issues.

3 Open-World Few-Shot Object Detection

3.1 Preliminary

Few-Shot Object Detection. We revisit few-shot object detection following previous works [11]. We first split the classes into base classes C_b with abundant instances and novel classes C_n with only K shot instances for each category, forming two sub-datasets named D_b and D_n, respectively. The goal of FSOD is to detect objects from novel classes C_n with only a few annotations. As a pioneer of the transfer-learning paradigm in FSOD, TFA proposes a two-stage fine-tuning approach that adopts Faster R-CNN as the base detector. In the first stage, TFA trains the feature extractor and the box predictor on base data D_b. Then in the second stage, TFA freezes the feature extractor and only fine-tunes the last classification and regression layers on a balanced dataset with objects from both base and novel classes. Despite the superior performance, its proposal generator usually overfits to trained classes and cannot adapt to the real world with unknown proposals. Therefore, a strong proposal generator supporting both known and unknown objects is needed for those detectors.

Class-Agnostic Object Detection aims to locate all instances in the image without classifying them. Unlike FSOD, CAOD only leverages a model trained on D_b to localize objects from both C_b and C_n, eliminating the fine-tuning stage. As the first attempt to explore the open-world proposals, OLN proposes an object localization network based on Faster R-CNN to recognize novel objects by learning pure localization cues. Compared to existing proposal generators, OLN replaces the classifiers in both RPN and RoI heads with localization quality estimators. Specifically, in the RPN module, centerness is chosen as the box localization quality target. In addition, the standard box-delta targets $(\mathrm{x}, \mathrm{y}, \mathrm{w}, \mathrm{h})$ are replaced with distances from the location to four sides of the ground-truth box (l, r, t, b) for the box regression. In the RoI head module, IoU is chosen as the box localization quality target to refine the proposal scoring and avoid over-fitting to the known classes. Although OLN can achieve strong performance on novel classes, the scalability is still limited for lack of multi-label information, i.e., C_b. Therefore, we propose a novel problem setting of open-world few-shot object detection based on few-shot learning, described as follows.

Our Problem Setting. In this section, we propose a new task, open-world few-shot object detection (OFOD), based on existing literature on class-agnostic object detection and few-shot object detection. For an object detection dataset D, we split it into three sub-datasets, a base set D_b with abundant annotated instances from C_b, a novel set D_n with a few examples of C_n, and an unknown set D_{uk} with a set of unknown classes but not labeled yet. Note that the three sets of classes are not overlapped, i.e., $C_b \cap C_n \cap C_{uk} = \emptyset$. Our goal is to learn a model to detect both C_b and C_n from the combination of D_b and D_n, and localize C_{uk} from D_{uk} without extra training samples. Compared with CAOD, our task additional needs to classify objects from C_b and make a good separation of C_n and C_{uk} by using a few annotations of C_n. Compared with FSOD, our task not only needs to detect objects from C_b and C_n, but also localize objects from C_{uk} (we do not need to classify them).

4 OFDet

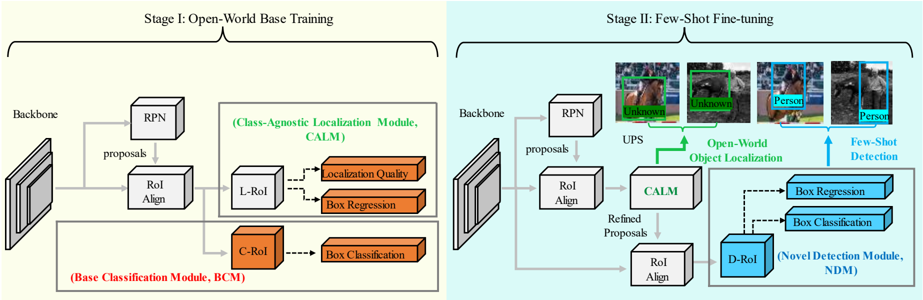

Fig. 3. The architecture of OFDet for open-world few-shot object detection. OFDet extends OLN [6] with additional fine-tuning stage. In the base training stage, Base Classification Module (BCM) is inserted into the framework for extracting the base features. In the fine-tuning stage, Novel Detection Module (NDM) is utilized to detect the novel classes. Class-Agnostic Localization Module (CALM) provides high-quality class-agnostic proposals for known object detection and unknown object localization. Besides, an Unknown Proposals Selection (UPS) algorithm is proposed to recognize unknown objects during inference. Moreover, the orange blocks, including C-RoI head and Box Classification in BCM and Box Regression in CALM, are pre-trained for the blue in NDM for fine-tuning.

As shown in **Fig. 3**, we propose a novel network called OFDet following the two-stage training approach, which consists of a backbone, Region Proposal Network

(RPN), Class-Agnostic Localization Module (CALM), Base Classification Module (BCM), and Novel Detection Module (NDM) based on the Faster R-CNN [2].

4.1 Stage I

During the base training stage, our architecture consists of the backbone, RPN, BCM and CALM. The box regressor in CALM and the box classifier in BCM are responsible for detecting base classes and localizing unknown objects. In this section, we will introduce BCM and CALM, respectively.

Class-Agnostic Localization Module. The Class-Agnostic Localization Module (CALM) is a variant of the R-CNN detection head, including a localization RoI head (L-RoI head), a bounding box regressor and a localization quality estimator [6]. Different from R-CNN, CALM does not distinguish the class label of a proposal, i.e., class-agnostic proposal generation. CALM is used to compute the four-dimensional coordinates of the proposals and corresponding localization quality. It plays an essential role in recognizing both known and unknown objects during two stages.

Base Classification Module. As discussed before, though the class-agnostic proposal generator has better localization ability for all foreground objects, it still lacks supervision of base labels. To obtain discriminative information in the first stage, a naive solution is to add another classifier following the L-RoI head of CALM. However, there may be a conflict between localization and classification tasks if both output layers follow the same RoI head, which could harm both the generalization ability for unknown classes and the classification ability for novel classes. Therefore, to avoid reducing the generalization ability to the unknown objects, we add an extra Base Classification Module (BCM) in the first stage to decouple the branches of localization and classification. BCM consists of the C-RoI head and the box classifier for base classes. We do not need to supervise the regression features since they are already obtained in CALM, so we can directly extract the classification feature for adapting to the base classes.

Open-World Base Training. In the first stage, the box regressor in CALM and box classifier in Base Classification Module are responsible for detecting base classes and localizing unknown objects. Thus, the limitation of no classification information in the original localization network can be compensated. The total loss in the first stage can be described as:

$$\mathcal{L}_{stage1} = \underbrace{(\lambda_1 \cdot \mathcal{L}_{loc} + \mathcal{L}_{reg})}_{RPN} + \underbrace{\mathcal{L}_{loc} + \mathcal{L}_{reg}}_{CALM} + \underbrace{\mathcal{L}_{cls}}_{BCM} \qquad (1)$$

where $\mathcal{L}_{cls}$ in BCM is a cross-entropy loss for the base classifier, and $\mathcal{L}_{reg}$ and $\mathcal{L}_{loc}$ in RPN and CALM are both L1 losses for the class-agnostic box regressor, and $\lambda_1 = 8$.

4.2 Stage II

In the second stage, the C-RoI head and its classifier layer in BCM are pre-trained for Novel Detection Module to detect novel classes with a few examples and then removed. In addition, an Unknown Proposal Selection is proposed for localizing the unknown objects during inference. We will introduce them next.

Novel Detection Module. As described in the task of OFOD, the detector needs to further operate on the localized unknown objects of interest in more applications. Therefore, we propose a Novel Detection Module (NDM) to detect novel objects from the unknown, which only have a few instances per class. As shown in **Fig. 3**, NDM is followed by the CALM and only applied in the fine-tuning stage, composed of a D-RoI head, a box regressor and a classifier. We use CALM to generate a set of proposals and then feed these proposals to NDM for further multi-class classification and regression. Since we have abundant instances of base classes and a few examples of novel classes, it is easy to learn a good separation of different classes and avoid training from scratch which is time-consuming and inefficient.

Unknown Proposal Selection. After we detect novel objects from NDM, we still need to localize the unknown objects which may be contained in the class-agnostic proposals by CALM. Therefore, we propose an Unknown Proposals Selection (UPS) algorithm that enables CALM to produce potential detections for unknown classes. As shown in Algorithm I, we first obtain the non-positive proposals and their corresponding objectness after CALM. As described in Line 3 to 8 of Algorithm I, we choose proposals with objectness scores greater than θ_1. After that, for each potential unknown proposals p_{uk}, we compute the IoU between p_{uk} and known proposals $\mathcal{P}_{uk}$ as $\mathcal{D}_p$ in Line 10, whose max IoU is computed as m_p in Line 11. If the IoUs between a potential unknown proposal and all known proposals are small, we select them as unknown proposals. More specifically, we select the proposals with an IoU less than θ_2 as unknown proposals (Line 12 to 13 in Algorithm I). Finally, we apply Non-Maximum Suppression (NMS) to filter out potential unknown proposals and then select top-k proposals as the final unknown predictions $\mathcal{P}_{uk}$ (Line 16). Then we concatenate unknown proposals and known proposals with their categories to evaluate both the average precision for known classes and classes-agnostic average recall for unknown objects, respectively.

Few-shot Fine-tuning. As shown in the second stage of **Fig. 3**, the weights of the D-RoI head are initialized by the weights of the C-RoI head, and the classification weights and the regression weights of base classes in NDM are initialized by the corresponding layers in BCM, respectively. The weights corresponding to the novel classes are randomly initialized. The network obtains high objectness scores of proposals and their corresponding deltas from the output of CALM. Then we take top-scoring k_1 proposals and perform RoI Align operation again to extract the refined region features for better localization. For detecting novel classes, we feed these features into D-

RoI head and finally carry out classification and regression tasks with two separate fully connected layers in NDM. The rest of the proposals which do not contain base or novel classes are utilized to localize the unknown objects with the aid of UPS. Following [12], we adopt a cosine similarity for box classifier. The total loss in the second stage can be described as:

$$\mathcal{L}_{stage2} = \underbrace{(\lambda_1 \cdot \mathcal{L}_{loc} + \mathcal{L}_{reg})}_{RPN} + \underbrace{\mathcal{L}_{loc} + \mathcal{L}_{reg}}_{CALM} + \underbrace{\mathcal{L}_{cls} + \mathcal{L}_{reg}}_{NDM} \tag{2}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ in NDM are a cross-entropy loss for box classification and a smoothed L1 loss for box regression, respectively.

```
Algorithm 1: Unknown Proposals Selection (UPS)
Input:
    $\mathcal{P}$ is a set of the proposals, $\mathcal{P}_{kn}$ is a set of known proposals, $\mathcal{P}_{kn} \subseteq \mathcal{P}$
    $\mathcal{S}$ is a set of the objectness of the proposals, $\mathcal{S}_{\mathrm{kn}}$ is a set of the objectness of the known proposals, $\mathcal{S} \subseteq \mathcal{S}_{kn}$
    $\theta_1$ is a hyperparameter of objectness threshold
    $\theta_2$ is a hyperparameter of IoU threshold
    $\mathcal{K}_{uk}$ is a hyperparameter of the number of unknown proposals
Output:
    $\mathcal{P}_{uk}$ is a set of unknown proposals
obtain non-positive proposals from NDM: $\mathcal{P}_{np} = \mathcal{P} - \mathcal{P}_{kn}$;
obtain the objectness of non-positive proposals: $\mathcal{S}_{np} = \mathcal{S} - \mathcal{S}_{kn}$;
for each non-positive proposal $p_{np} \in \mathcal{P}_{np}$ and corresponding objectness $s_{np} \in \mathcal{S}_{np}$ do
    build an empty set to choose proposals with the objectness scores greater than $\theta_1$: $\mathcal{P}_{fo} \leftarrow \emptyset$;
    if $s_{np} \geq \theta_1$ then
        $\mathcal{P}_{fo} = \mathcal{P}_{fo} \cup p_{np}$;
    end if
end for
for each proposal $p_{uk} \in \mathcal{P}_{fo}$ do
    compute IoU between $p_{uk}$ and $\mathcal{P}_{kn}$; $\mathcal{D}_p = IoU(\mathcal{P}_{kn}, p_{uk})$;
    compute max of $\mathcal{D}_p$: $m_p = Max(\mathcal{D}_p)$;
    if $m_p \leq \theta_2$ then
        $\mathcal{P}_{uk} = \mathcal{P}_{uk} \cup p_{uk}$;
    end if
end for
$P_{uk} \leftarrow$ select top $\mathcal{K}_{uk}$ proposals after NMS operation from $\mathcal{P}_{uk}$;
return $\mathcal{P}_{uk}$;
```

4.3 Inference

During inference, the final objectness score s of a proposal in CALM is computed as the mean of RPN localization quality r and CALM localization quality c, i.e., $s = \sqrt{(r \cdot c)}$. In the second stage, we take top-scoring k_2 proposals with NMS threshold of θ_3 obtained by CALM to feed into NDM for classification and regression, where

k_1, k_2 are both 5000 and θ_3 is set as 0.9. For recognizing the unknown objects, $\theta_1 = 0.4$, $\theta_2 = 1.0$, $K_{uk} = 20$ of UPS are set for recalling objects.

5 Experiments

5.1 Experiments Setting

Existing benchmarks. We evaluate our method for several tasks on the widely used detection benchmarks, PASCAL VOC and MS-COCO. For FSOD, we follow previous works [11, 12] and use the same class and data splits for a fair comparison. For PASCAL VOC, there are 20 classes randomly grouped into three splits and divided into 15 base classes and 5 novel classes. For few-shot fine-tuning, each of the novel classes has $K = 1,2,3,5,10$ instances from the combination of the trainval sets of PASCAL VOC 07 and 12. PASCAL VOC 2007 test set is used for evaluation. For COCO, 60 classes that are not overlapped with VOC classes are selected as base classes and we keep the remaining 20 classes as the novel classes with 5,10,30 instances. We use 5000 images from the validation set for evaluation and the rest for training. For CAOD, we have two splits. The first split follows the setting of OLN [6], where we use only box annotations of 20 classes for training and 60 unseen classes for testing. The second split follows the setting of FSOD instead, where we use only box annotations of 60 classes for training and the rest for testing. For **open-world few-shot object detection,** we divide COCO into 60 base classes, 15 novel classes, and 5 unknown classes with three random groups. The unknown categories of each unknown set are identical to the ones on each novel set of PASCAL VOC (e.g., motorcycle, bus, bird, cow and coach in Novel Set 1).

Evalutaion. We introduce the evaluation setting for three tasks, respectively. For CAOD experiments, following [6], class-agnostic COCO-style AR over N proposals on the novel classes are used for both COCO and PASCAL VOC. For FSOD experiments, we measure COCO-style mAP and AP_{50} for COCO and PASCAL VOC, respectively. For OFOD experiments, we together measure COCO-style mAP and unknown AR over N unknown proposals for both known classes and unknown classes, respectively.

Implementation details. We implement our method based on detectron2 [34]. The SGD optimizer with momentum 0.9 and weight decay 1e-4 is utilized to optimize our network over 8 GPUs with 16 images per mini-batch (2 images per GPU). We use ResNet-101 [35] pre-trained on ImageNet [36] as the backbone for most experiments, except for COCO's split 1 of class-agnostic object detection experiments, in which ResNet-50 is used to align with [6] for a fair comparison. The learning rate is set to be 0.02 during base training and 0.01 during fine-tuning. Inspired by [28], we decouple the backward gradient from RPN and several RCNN modules in the first stage. For PASCAL VOC, we freeze most layers and only fine-tune NDM for better performance. In addition, random-lighting augmentation is applied when fine-tuning our detector on the few-shot datasets, where the lighting scale is set to be 0.8. For other hyperparameters, we follow the settings of OLN and TFA.

5.2 Results on Class-Agnostic Object Detection

Table 2. Experimental results of CAOD on COCO. We evaluate our method performance (class-agnostic AR) on two splits under 10, 30-shot settings. ∗: reproduced by us.

Method / Split	AR10		AR50		AR100	
	Split 1	Split 2	Split 1	Split 2	Split 1	Split 2
OLN [6]	15.9*	10.0*	25.5*	16.2*	29.6*	18.4*
OFDet-10shot(ours)	18.2	17.6	27.9	25.1	31.7	27.7
OFDet-30shot(ours)	21.5	21.8	32.2	30.4	37.0	33.3

For a fair comparison, we reproduce the unknown classes results of OLN based on detectron2. We present our evaluation results of COCO on two different splits in **Table 2**. It can be seen that our method can significantly improve the recall of novel classes compared to OLN. The AR10 of split 1 in our method can achieve 18.2% and 21.5% under 10-shot and 30-shot settings, boosting OLN by 2.3% and 5.6%. For split 2, our work can also achieve 17.6% and 21.8% in AR10 evaluation, which gains 7.6% and 11.8\% AR compared to OLN. These results indicate that we can largely improve AR by adding a few annotations for those classes unseen in the previous stage.

5.3 Results on Few-Shot Object Detection

PASCAL VOC. We provide the results of three splits for PASCAL VOC in **Table 3**. We would like to emphasize that recent methods are not shown in the table because we select TFA as the baseline and focus on the more challenging OFOD task. We make a detailed comparison with TFA to verify the effectiveness of our work. As shown in **Table 3**, our method is superior to the transfer-learning baseline TFA. To be specific, our method outperforms TFA by 3.2%, 4.9%, 4.9%, 8.4%, 3.2% and 3.9%, 0.7%, 3.9%, 4.0%, 5.6% and 4.8%, 9.0%, 0.8%, 1.9%, 4.7% for $K = 1,2,3,5,10$ on Novel Set 1, Set 2 and Set 3. Besides, our method can outperform meta-learning approaches [11, 24, 38] under most settings.

COCO. Table 4 demonstrates the few-shot detection results for COCO. Our methods can achieve 9.6%, 12.7% and 17.5% in 5-shot, 10-shot and 30-shot settings, which gains 2.2%, 2.7% and 3.8% for $K = 5, 10, 30$ shots per category than TFA. Meanwhile, our method can also outperform meta-learning baseline [2416] under most settings, showing our methods' strong robustness and generalization ability in a more realistic and complex scenario such as COCO.

Table 3. Experimental results of FSOD on PASCAL VOC. We evaluate our method (AP_{50}) under 1, 2, 3, 5, 10 shots settings. The best results and the second best results are colored in red and blue, respectively, the same below.

Method / Shots	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLO-FT	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FRCN-FT	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
LSTD	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
FSRW	14.8	15.5	26.7	33.9	47.2	15.7	15.2	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet[37]	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
MetaR-CNN	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
OFDet(ours)	42.0	41.0	49.6	56.7	59.2	27.4	27.6	38.0	39.1	44.7	35.6	43.8	43.6	51.4	54.5

Table 4. Experiment results of FSOD on COCO. We evaluate our method (mAP) under 5, 10, 30 shots settings.

Method/Shots	5	10	30
FRCN-FT	4.0	6.5	11.1
FSRW	-	5.6	9.1
MetaDet	-	7.1	11.3
Meta R-CNN	-	8.7	12.4
TFA	7.7	10.0	13.7
MPSR [39]	8.7	9.8	14.1
FSDetView [38]	10.7	12.5	14.7
OFDet(ours)	9.6	12.7	17.5

5.4 Results on OFOD

We conduct open-world few-shot object detection experiments on COCO. We compare OFDet with several baselines, including FRCN-ALL [2], TFA [12] and Meta R-CNN [24]. We train these models until convergence to obtain robust performance under OFOD setting. As shown in **Table 5**, though these detectors can detect novel classes with a few samples, they fail to localize the unknown objects under 10-shot or 30-shot setting. The reason may be that their classifier cannot recognize the unknown objects, harming the unknown recall evaluation. To further check the effectiveness of our method, we additional introduce these methods with our proposed UPS to select the unknown proposals from RPN. In contrast, the unknown AR of these methods can be largely improved for three unknown sets under 10-shot setting and in 30-shot setting by using UPS. Finally, compared with the formal results, our method can achieve state-of-the-art performance on unknown AR with 30.2%, 30.1% and 24.4% under 10-shot setting, 32.5%, 31.2% and 24.5% in 30-shot setting, respectively, while keeping the competitive performance of detecting novel classes. The results in **Table 5** reveal that our method has a better balance between localizing unknown objects and detecting novel classes.

Table 5. Experimental results of OFOD on COCO. We report novel AP (nAP) and unknown AR (uAR) on three splits. FRCN-ALL stands for fine-tuning entire Faster R-CNN without the freeze strategy. The term w/ UPS indicates the existing method with proposed UPS to select unknown proposals. All results are averages of multiple random runs.

Method / Split 1	Unk Set1		Unk Set 2		Unk Set 3	
	nAP	uAR	nAP	uAR	nAP	uAR
10-shot						
FRCN-ALL	8.3	0.0	9.4	0	10.5	0
FRCN-ALL w/ UPS	8.3	19.5	9.4	13.9	10.5	15.5
Meta R-CNN	8.6	0.0	9.0	0.0	10.8	0.0
Meta R-CNN w/ UPS	8.6	15.9	9.0	15.4	10.8	14.1
TFA	9.3	0.0	9.5	0.0	11.2	0.0
TFA w/ UPS	9.3	13.6	9.5	12.4	11.2	10.6
OFDet(ours)	11.1	30.2	11.2	30.1	12.6	24.4
30-shot						
FRCN-ALL	13.5	0.0	14.6	0.0	14.7	0.0
FRCN-ALL w/ UPS	13.5	22.4	14.6	17.6	14.7	20.1
Meta R-CNN	11.0	0.0	11.1	0.0	12.3	0.0
Meta R-CNN w/ UPS	11.0	14.4	11.1	14.8	12.3	13.7
TFA	12.8	0.0	13.6	0.0	14.5	0.0
TFA w/ UPS	12.8	20.0	13.6	18.0	14.5	17.1
OFDet(ours)	15.7	32.5	16.7	31.2	17.1	24.5

5.5 Ablation Study

Effectiveness of UPS. As a plug-and-play method, the proposed UPS is easily utilized for two-stage object detectors to help the existing proposal generators localize more unknown proposals. Here, we apply UPS to other previous approaches, including FRCN-ALL, TFA and Meta R-CNN. As shown in **Table 5** and discussed in Sec.5.4 before, we obverse that using UPS can achieve higher performance, with about 100% relative improvement in Unk AR, demonstrating the effectiveness of our UPS.

Ablation of Weight Initialization in NDM. We conduct more ablations on 10-shot split 1 of PASCAL VOC to carefully analyze how the weight initialization in NDM contributes to the performance of novel classes. All results are shown in Table 6. Specifically, the first row demonstrates that base model without BCM only achieves 30.2% nAP, indicating that the lack of discriminative information of different classes leads our method to generalize poorly to novel classes. Next, we take four progressive steps to explore the fine-tuning strategy used in NDM: (1) use the weights of the C-RoI head. We find that fine-tuning D-RoI can achieve 55.5% with a large margin. (2) add the regression weights in CALM under the setting of (1). The result also makes a stable boost with 2.1%. (3) add the classification weights in BCM based on (1). This improvement further gains 2.8% on novel classes. Finally, we integrate the above three pre-trained weights into the original OFDet, which makes a significant boost with 28.5% for 10-shot, proving the effectiveness of BCM and fine-tuning strategy to alleviate overfitting on base classes.

Table 6**.** Ablation of weight initialization of NDM. The D-RoI head, Cls Layer and Reg Layer indicate that we initialize these modules with the C-RoI head in BCM, the classification layer in BCM, and the regression layer in CALM, respectively.

Method	D-RoI Head	Cls Layer	Reg Layer	nAP
OFDet				30.2
	√			55.5
	√		√	57.6
	√	√		58.3
	√	√	√	59.2

Visualization. 错误!未找到引用源。 is our qualitative results on COCO. Our method can not only detect both base and novel objects, but also localize unknown objects without training samples.

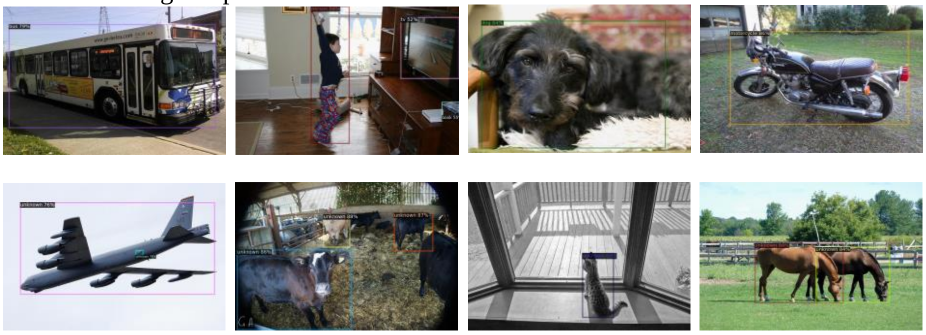

Fig. 4. Visualization results of our proposed method on 10-shot of COCO dataset. Top-row: novel objects. Bottom-row: unknown objects.

6 Conclusion

General object detection lacks the ability of generalization to unknown objects even though this task has made significant progress. In this work, we propose Open-World Few-Shot Object Detection, where our proposed OFDet extends a class-agnostic object detector based on the two-stage approach to perform detection for novel classes with a few available examples and localization for unknown class with an Unknown Proposals Selection in more realistic scenarios. Extensive experiments on the challenging benchmarks PASCAL VOC and MS-COCO illustrate that our proposed OFDet can perform well on CAOD, FSOD and our proposed OFOD. We hope our work will inspire the vision community to further explore this novel task for practical applications.

References

1. Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
2. Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).
3. Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.
4. Alexe, Bogdan, Thomas Deselaers, and Vittorio Ferrari. "Measuring the objectness of image windows." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2189-2202.
5. Uijlings, Jasper RR, et al. "Selective search for object recognition." International journal of computer vision 104 (2013): 154-171.
6. Kim D, Lin T Y, Angelova A, et al. Learning open-world object proposals without learning to classify[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5453-5460.
7. Scheirer, Walter J., et al. "Toward open set recognition." IEEE transactions on pattern analysis and machine intelligence 35.7 (2012): 1757-1772.
8. Bendale, Abhijit, and Terrance Boult. "Towards open world recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
9. Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
10. Chen, Hao, et al. "Lstd: A low-shot transfer detector for object detection." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
11. Kang, Bingyi, et al. "Few-shot object detection via feature reweighting." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
12. Wang, Xin, et al. "Frustratingly Simple Few-Shot Object Detection." International Conference on Machine Learning. PMLR, 2020.
13. Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
14. Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88 (2009): 303-308.

15. Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.
16. Liu, Wei, et al. "Ssd: Single shot multibox detector." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.
17. Farhadi, Ali, and Joseph Redmon. "Yolov3: An incremental improvement." Computer vision and pattern recognition. Vol. 1804. Berlin/Heidelberg, Germany: Springer, 2018.
18. Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.
19. He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
20. Carion, Nicolas, et al. "End-to-end object detection with transformers." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020.
21. Rosenfeld, Azriel, and Mark Thurston. "Edge and curve detection for visual scene analysis." IEEE Transactions on computers 100.5 (1971): 562-569.
22. Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in neural information processing systems 30 (2017).
23. Vinyals, Oriol, et al. "Matching networks for one shot learning." Advances in neural information processing systems 29 (2016).
24. Yan, Xiaopeng, et al. "Meta r-cnn: Towards general solver for instance-level low-shot learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
25. Hu, Hanzhe, et al. "Dense relation distillation with context-aware aggregation for few-shot object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
26. Fan, Zhibo, et al. "Generalized few-shot object detection without forgetting." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
27. Sun, Bo, et al. "Fsce: Few-shot object detection via contrastive proposal encoding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
28. Qiao, Limeng, et al. "Defrcn: Decoupled faster r-cnn for few-shot object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
29. Endres, Ian, and Derek Hoiem. "Category-independent object proposals with diverse ranking." IEEE transactions on pattern analysis and machine intelligence 36.2 (2013): 222-234.
30. Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.
31. Cai, Zhaowei, and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
32. Vu, Thang, et al. "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution." Advances in neural information processing systems 32 (2019).
33. Wang, Jiaqi, et al. "Region proposal by guided anchoring." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
34. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

35. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
36. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60.6 (2017): 84-90.
37. Wang, Yu-Xiong, Deva Ramanan, and Martial Hebert. "Meta-learning to detect rare objects." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
38. Xiao, Yang, Vincent Lepetit, and Renaud Marlet. "Few-shot object detection and viewpoint estimation for objects in the wild." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.3 (2022): 3090-3106.
39. Wu, Jiaxi, et al. "Multi-scale positive sample refinement for few-shot object detection." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer International Publishing, 2020.
40. Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
41. Sun, Peize, et al. "Sparse r-cnn: End-to-end object detection with learnable proposals." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
42. Fan, Qi, et al. "Few-shot object detection with attention-RPN and multi-relation detector." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.