Exploring Base-Class Suppression with Prior Guidance for Bias-Free One-Shot Object Detection

Wenwen Zhang[1], Yun Hu[2], Hangguan Shan[1], Eryun Liu[1]*

[1]College of Information Science and Electronic Engineering, Zhejiang University, China
[2]School of Information Science and Technology, ShanghaiTech University, China
{wenwenzhang, hshan, eryunliu}@zju.edu.cn, huyun@shanghaitech.edu.cn

Abstract

One-shot object detection (OSOD) aims to detect all object instances towards the given category specified by a query image. Most existing studies in OSOD endeavor to establish effective cross-image correlation with limited query information, however, ignoring the problems of the model bias towards the base classes and the generalization degradation on the novel classes. Observing this, we propose a novel algorithm, namely Base-class Suppression with Prior Guidance (BSPG) network to achieve bias-free OSOD. Specifically, the objects of base categories can be detected by a base-class predictor and eliminated by a base-class suppression module (BcS). Moreover, a prior guidance module (PG) is designed to calculate the correlation of high-level features in a non-parametric manner, producing a class-agnostic prior map with unbiased semantic information to guide the subsequent detection process. Equipped with the proposed two modules, we endow the model with a strong discriminative ability to distinguish the target objects from distractors belonging to the base classes. Extensive experiments show that our method outperforms the previous techniques by a large margin and achieves new state-of-the-art performance under various evaluation settings.

Introduction

Benefiting from the flourishing of deep convolutional neural networks, object detection has made tremendous progress over the past few years (Ren et al. 2015; Redmon et al. 2016; He et al. 2017). Nevertheless, most of the advanced methods heavily rely on large-scale labeled datasets (Deng et al. 2009; Lin et al. 2014), and they may struggle with new applications where novel-class objects are not witnessed during the training phase. In light of the powerful cognitive ability of humans to recognize new objects with only a few examples, few-shot learning (FSL) has emerged as a promising technique (Vinyals et al. 2016; Sung et al. 2018). FSL constructs models that can generalize to new classes with limited annotated data, offering a potential solution for object detection in scenarios involving novel-class objects.

In this paper, we undertake the application of FSL in the field of object detection, termed one-shot object detection (OSOD), where the model aims to detect all instances of the given category specified by only one query image patch. In recent years, the field of few-shot object detection has thrived (Fan et al. 2020a; Chen et al. 2021; Han et al. 2022), in which prevalent approaches often incorporate transfer-learning, meta-learning, and metric-learning to deal with the task. Most existing works in OSOD adopt the metric-learning paradigm and recognize new objects based on the similarity metrics between image pairs without fine-tuning (Hsieh et al. 2019; Zhang et al. 2022b; Yang et al. 2022). However, they are generally dedicated to exploring effective cross-image feature correlation to better use the limited information, neglecting the phenomenon of the model bias towards the base classes and the generalization degradation on the novel classes (Fan et al. 2021; Lang et al. 2022). Due to the extremely unbalanced distribution of base and novel-class datasets, the learned model will inevitably

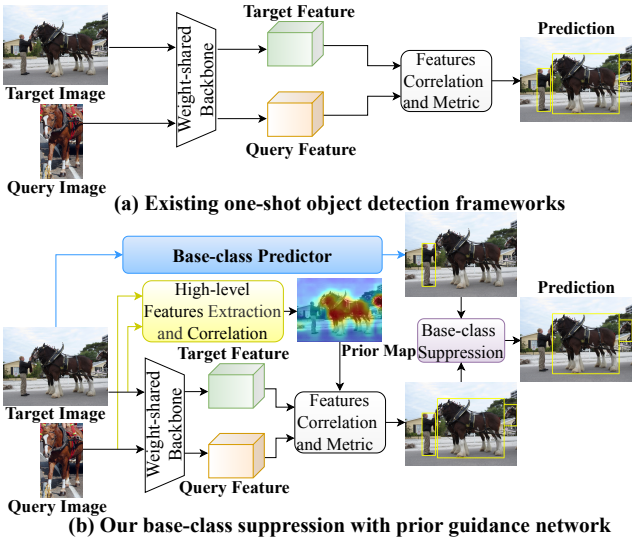

Figure 1: Comparison between existing frameworks and ours. Existing models mostly exhibit a preference for familiar base class (e.g., *person*) rather than novel objects specified by the query image (e.g., *horse*). We propose a base-class suppression module to eliminate base-class distractors, and a non-parametric prior guidance module to generate a class-agnostic prior map to guide the detection process.

*Corresponding author.

fit the feature distribution of the abundant training data and tend to exhibit a preference for the base classes over the given novel category. As illustrated in Fig. 1(a), the conventional OSOD model easily suffers from false positive detections on the base-class objects, leading to a decrease in performance for novel categories.

To address the aforementioned problems, we tackle the OSOD task from a new perspective, improving the model by suppressing the distractors belonging to base classes. Specifically, a complementary branch named base-class predictor is introduced to detect the objects of base classes, which is pre-trained on the base-class dataset following a traditional paradigm. Then, we rectify the coarse detection results derived from the general OSOD network (novel-class predictor) by Base-class Suppression (BcS) module. As shown in Fig. 1(b), equipped with the BcS module, the falsely detected objects are obviously suppressed, thus realizing accurate detection for the specified novel category and explicitly mitigating the model bias problem.

Rather than pursuing well-designed interaction, which may involve numerous learnable parameters and potentially lead to a bias towards the base classes, we propose a non-parametric prior guidance module to facilitate the recognition of novel classes. In this module, we calculate the semantic relation between high-level features to generate a class-agnostic prior map, in which the regions of target features belonging to the co-existing objects can be activated, as illustrated in Fig. 1(b). The prior map with unbiased semantic cues can guide the subsequent detection procedure and help the model distinguish target objects from the background. Since the prior generation is non-parametric, the model can learn more general patterns and retain generalization ability, thereby implicitly alleviating the bias problem. By integrating these two proposed modules, we establish a novel OSOD framework, termed Base-class Suppression with Prior Guidance (BSPG) network. These two components complement each other to promote bias-free OSOD and enhance generalization for novel classes. In summary, our main contributions can be summarized as follows:

- We propose a BSPG network to address the model bias towards base classes problem, which has been neglected in previous works on OSOD.
- We introduce a base-class predictor to detect the objects of base classes, and a base-class suppression module to eliminate them, facilitating the detection of novel objects.
- We design a non-parametric prior guidance module to generate a class-agnostic prior map with unbiased semantic cues and guide the detection procedure.
- Extensive experiments illustrate that our proposed approach yields new state-of-the-art results, which validate its effectiveness and robustness.

Related Work

General Object Detection. Object detection aims to locate objects of seen classes and assign a category label to each object instance. General detectors can be broadly divided into two streams: one-stage methods (Liu et al. 2016; Redmon and Farhadi 2017; Duan et al. 2019) and two-stage ones (Girshick 2015; Ren et al. 2015). Currently, OSOD is still in its early stage of research. To pursue a high algorithm accuracy, our model is developed based on the two-stage detector of Faster R-CNN (Ren et al. 2015).

Few-Shot Object Detection. Few-shot object detection (FSOD) performs object detection on a target image conditioned on a limited number of query images. Existing FSOD methods can be generally categorized into three directions: transfer learning, meta learning, and metric learning methods. The transfer-learning based model is initially pre-trained on abundant base data and subsequently fine-tuned on a limited set of novel examples. DeFRCN (Qiao et al. 2021) performs decoupling among multiple modules of Faster R-CNN to boost performance. The meta-learning based methods are dedicated to learning efficient meta knowledge and fostering adaptation to novel categories. Meta R-CNN (Yan et al. 2019) extends Faster R-CNN by applying channel-wise attention to reweight the RoI features. The metric-learning based methods focus on exploring cross-image correlations to directly detect novel objects without fine-tuning. BHRL (Yang et al. 2022) proposes a multi-level relation module to establish semantic relations. OSOD, as an extreme case of FSOD, involves the localization and classification for novel objects with only one sample. Recent researches (Sun et al. 2021; Yang et al. 2021) suggest that the box regressor is capable of accurately localizing objects. Owing to the seriously unbalanced data distribution, the primary source of generalization degradation is misclassifying the instances of base classes as objects of interest. Therefore, we employ a base-class predictor to explicitly detect the base-class objects and further suppress them, and perform prior guidance in a non-parametric manner to generate unbiased prior knowledge.

Method

Problem Definition

Following the common configurations in previous literature (Yang et al. 2022), the object classes of the dataset are partitioned into two disjoint parts $\mathcal{C}_b \cap \mathcal{C}_n = \emptyset$. Here, $\mathcal{C}_b$ denotes base classes with abundant annotated data for training, and $\mathcal{C}_n$ represents novel classes with only one instance per category. We train our model with the episodic paradigm, where each episode contains a target-query pair. The one-shot object detector is expected to identify all instances of the same category specified by the query image q in the target image I. The model is continuously optimized using numerous available data of base classes $\mathcal{C}_b$ during the training phase. Once the training is completed, the detector can directly predict the objects of novel classes $\mathcal{C}_n$ in the target image conditioned on one query image without fine-tuning.

Overview

Fig. 2 sketches the overall architecture of our BSPG network. It comprises four essential components: a base-class predictor, a novel-class predictor, a base-class suppression, and a prior guidance module. We apply the two-stage training strategy to train the base and novel predictors, respectively. In the first stage, we optimize the base-class predic-

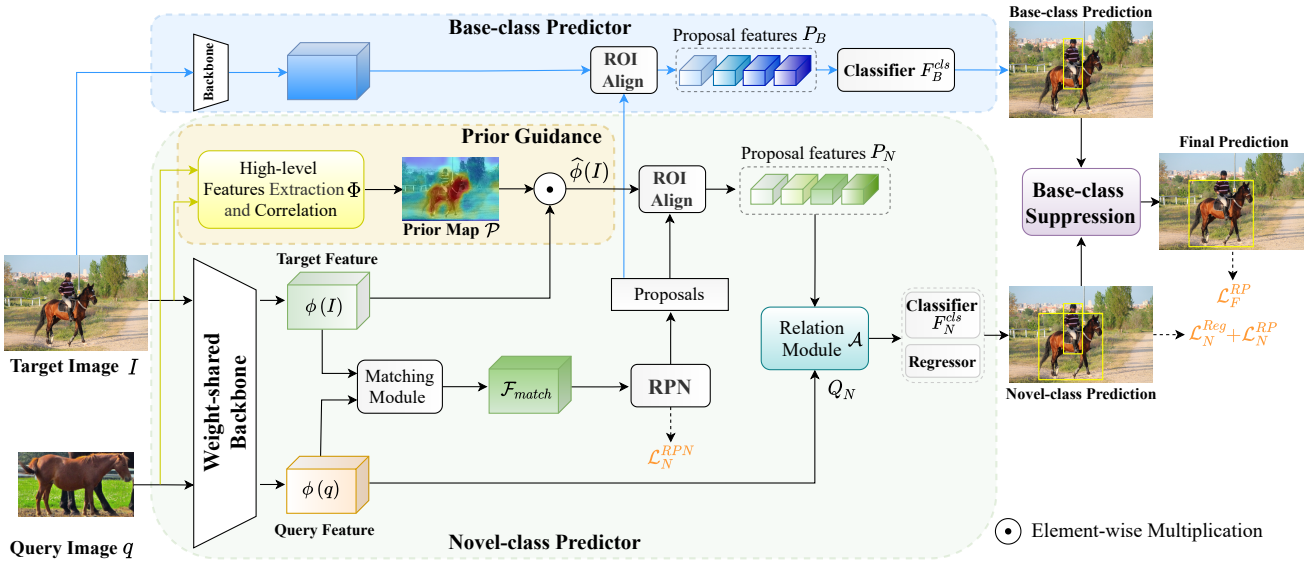

Figure 2: The overall architecture of the proposed BSPG, which contains four key components: a base-class predictor, a novel-class predictor, a base-class suppression, and a prior guidance module. Based on the base-class predictions, we refine the coarse novel-class predictions by base-class suppression to yield the final prediction results. Meanwhile, the prior guidance module generates a class-agnostic prior map, providing unbiased semantic cues to effectively guide the subsequent detection procedure.

tor on the base-class dataset following the traditional Faster R-CNN paradigm. For the second stage, we fix the parameters of the base predictor and only optimize the novel predictor and the BcS module. Concretely, the two predictors are deployed to respectively identify the objects of base and novel classes. Then, we rectify the coarse detection results obtained from the novel predictor by base-class suppression to yield the final results. Meanwhile, the non-parametric PG module calculates the cross-image correlation between high-level features and generates a class-agnostic prior map. The prior map serves as an indicator, providing unbiased semantic hints to guide the detection process effectively.

Base-class Suppression

Base-class Predictor. To alleviate the performance degradation caused by misclassifying base-class instances as target objects, we introduce a base-class predictor. This branch is specifically designed to explicitly detect objects belonging to the base categories. However, we noticed that the proposals generated by the base-class predictor are mostly active in the base categories, which may burden the following classification task for novel-class detection. For simplicity, we only focus on handling misclassification produced by the novel-class predictor. Therefore, the RPN of the base predictor is removed in the second training phase, and the proposals generated by the novel predictor are delivered to the base predictor for further proposal features retrieval and classification. The base-class classifier F_B^{cls} yields the base-class classification results for proposal features P_B:

$$Y_B = F_B^{cls}(P_B) \in \mathbb{R}^{K\times(1+S)}, \tag{1}$$

where K is the number of proposals, S denotes the number of base categories (equaling 60 for COCO dataset (Lin et al. 2014), and equaling 16 for VOC dataset (Everingham et al. 2010) under 1-shot setting), and Y_B denotes the base-class prediction probability over S categories.

Novel-class Predictor. Given a query image q and a target image I, the novel-class predictor aims to detect the object instances in I towards the given category specified by q. Specifically, the features $\phi(q)$, $\phi(I)$ extracted by the siamese ResNet-50 (He et al. 2016) with feature pyramid network (FPN) (Lin et al. 2017) are fed into the matching module (Michaelis et al. 2018), intending to accomplish feature matching.

$$\mathcal{F}_{diff} = |\phi(I) - GAP(\phi(q))|, \tag{2}$$

$$\mathcal{F}_{match} = Conv([\mathcal{F}_{diff}, \phi(I)]), \tag{3}$$

where $|\cdot|$ denotes the absolute value operator, GAP denotes global average pooling operation, $Conv$ denotes convolutional layers, $\mathcal{F}_{diff}$ denotes pointwise L1 difference, and $[\cdot,\cdot]$ denotes concatenation operation. By propagating the discriminative similarity feature $\mathcal{F}_{match}$ to RPN, it can generate more expected proposals with high potential including target objects, and filter out negative objects not belonging to the query category. Then we use the RoI Align layer to obtain proposal features P_N based on the proposals.

To distinguish whether the proposals belong to the target category or not, we follow BHRL (Yang et al. 2022) to introduce a hierarchical relation module $\mathcal{A}$ consisting of contrastive-level, salient-level, and attention-level correlations. The goal is to comprehensively measure the semantic relation and re-score proposals. Specifically, for contrastive-level relation $\mathcal{A}_c$, we compute the absolute difference between the query vector and each position of the proposal feature P_N. For salient-level relation $\mathcal{A}_s$, the query feature

Q_N is regarded as a convolution kernel to generate the relevant features with proposal feature P_N in a depth-wise manner (Fan et al. 2020a). For attention-level one $\mathcal{A}_a$, we calculate spatial attention matrix W_s to reweight the query feature and further compute the absolute difference.

$$\mathcal{A}(P_N, Q_N) = [\mathcal{A}_c(P_N, Q_N), \mathcal{A}_s(P_N, Q_N), \mathcal{A}_a(P_N, Q_N), P_N], \quad (4)$$

$$\mathcal{A}_c(P_N, Q_N) = Conv(|GAP(Q_N) - P_N|), \quad (5)$$
$$\mathcal{A}_s(P_N, Q_N) = Conv(GAP(Q_N) \otimes P_N), \quad (6)$$
$$\mathcal{A}_a(P_N, Q_N) = Conv(|W_s Q_N - P_N|), \quad (7)$$

$$W_s = softmax((Conv(P_N))^T Conv(Q_N)), \quad (8)$$

where $\otimes$ represents the features correlation in a depth-wise manner. Subsequently, we deliver the relation features to the novel-class classifier F_N^{cls} and obtain the coarse novel-class classification results Y_N:

$$Y_N = F_N^{cls}(\mathcal{A}(P_N, Q_N)) \in \mathbb{R}^{K\times 2}. \quad (9)$$

Base-class Suppression. The BcS module is designed to eliminate the false predictions on the base classes. For each proposal, if the proposal whose highest base-class prediction probability over S categories (excluding the background) is more than threshold α, base-class result $\mathcal{B}$ equals the corresponding highest base-class prediction score, otherwise 0.

$$\mathcal{B} = \begin{cases} \max\limits_{i\in\{1,2,\ldots,S\}}(Y_B^i) & \max\limits_{i\in\{1,2,\ldots,S\}}(Y_B^i) > \alpha \\ 0 & \text{otherwise} \end{cases}, \quad (10)$$

where α is set to 0.3 for COCO dataset and 0.7 for VOC dataset. Finally, the coarse detection results Y_N derived from the novel-class predictor are rectified by the base-class suppression, resulting in the final results Y_F:

$$Y_F = [Y_F^f, Y_F^b] = [\Psi_f(Y_N^f, \mathcal{B}), \Psi_b(Y_N^b, \mathcal{B})], \quad (11)$$
$$Y_F^f = W_N^f * Y_N^f + W_B^f * \mathcal{B}, \quad (12)$$
$$Y_F^b = W_N^b * Y_N^b + W_B^b * \mathcal{B}, \quad (13)$$

where Y_N^f and Y_N^b denote the foreground and background probability predicted by the novel-class predictor, respectively. Ψ_f and Ψ_b are learnable fully connected layers, which are deployed to refine the coarse novel-class scores by suppressing false predictions on the base-class objects. As expected in our experiment, when $\mathcal{B}$ is greater than 0, the learnable weight W_N^f is positive and W_B^f is negative to decrease the final foreground score Y_F^f. Conversely, the weight W_N^b is positive and W_B^b is also positive to increase the background score, aiming to suppress base-class objects. The BcS module adaptively learns to assign the weight to the base and novel-class prediction results. The higher the base-class confidence score, the more obviously it is suppressed.

Loss Function. For the second stage, the overall loss for training our model can be expressed by:

$$\mathcal{L} = \mathcal{L}_N^{RPN} + \mathcal{L}_N^{ROI}, \quad (14)$$

$$\mathcal{L}_N^{ROI} = \mathcal{L}_N^{Reg} + \lambda_1 \mathcal{L}_N^{RP}(Y_N, G) + \lambda_2 \mathcal{L}_F^{RP}(Y_F, G), \quad (15)$$

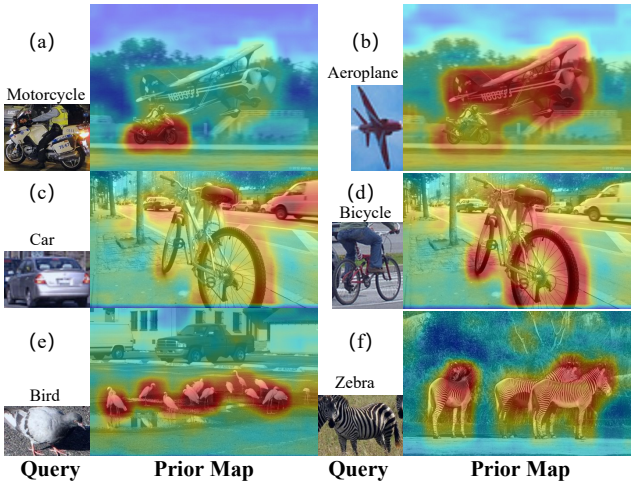

Figure 3: Visualizations of the prior map generated by our prior guidance module via high-level features correlation. We show the query image and prior map from left to right.

where $\mathcal{L}_N^{RPN}$ is the RPN loss of Faster-RCNN, and $\mathcal{L}_N^{ROI}$ contains the classification and regression losses in the ROI head of the novel predictor. $\mathcal{L}_N^{Reg}$ is the regression loss, $\mathcal{L}_N^{RP}$ and $\mathcal{L}_F^{RP}$ are the ratio-preserving losses defined in BHRL (Yang et al. 2022) for evaluating the coarse novel-class and final classification results, respectively. G denotes the ground-truth label. λ_1 and λ_2 are set to 0.5.

Prior Guidance

We observe that the well-designed modules with many learnable parameters in current prevalent OSOD models may inevitably introduce a bias towards base classes, hindering the accurate detection of novel classes. Inspired by (Tian et al. 2022), we propose a non-parametric prior guidance module to mine semantic correlations without sacrificing generalization ability. Concretely, our network adopts the ResNet-50 (He et al. 2016) as a frozen backbone Φ to encode high-level features $\Phi(q)$ and $\Phi(I)$ from the raw image-pair, where the parameters are pre-trained on the reduced ImageNet following the OSOD settings (Yang et al. 2022). Next, we calculate the element-wise relation map $\mathcal{R} \in \mathbb{R}^{H_I^h W_I^h \times H_q^h W_q^h}$ between $\Phi(I) \in \mathbb{R}^{C^h \times H_I^h \times W_I^h}$ and $\Phi(q) \in \mathbb{R}^{C^h \times H_q^h \times W_q^h}$ using cosine similarity function:

$$\mathcal{R} = \Theta(\Phi(I))^T \Theta(\Phi(q)), \quad (16)$$

where Θ denotes the L2 normalization. For each element in the target feature $\Phi(I)$, we select the maximum similarity among all elements of the query feature as the relation values to generate the prior map $\mathcal{P} \in \mathbb{R}^{H_I^h W_I^h \times 1}$:

$$\mathcal{P} = \max_{j\in\{1,2,\ldots,H_q^h W_q^h\}} \mathcal{R}(i, j). \quad (17)$$

A high activation value in $\mathcal{P}$ for an element of the target feature indicates that this element has an intense correlation with at least one element in the query feature (Tian et al.

Method	Base class					Novel class				
	split-1	split-2	split-3	split-4	Average	split-1	split-2	split-3	split-4	Average
CoAE (Hsieh et al. 2019)	42.2	40.2	39.9	41.3	40.9	23.4	23.6	20.5	20.4	22.0
AIT (Chen, Hsieh, and Liu 2021)	50.1	47.2	45.8	46.9	47.5	26.0	26.4	22.3	22.6	24.3
SaFT (Zhao, Guo, and Lu 2022)	49.2	47.2	47.9	49.0	48.3	27.8	27.6	21.0	23.0	24.9
BHRL (Yang et al. 2022)	56.0	52.1	52.6	53.4	53.5	26.1	29.0	22.7	24.5	25.6
BHRL* (Yang et al. 2022)	56.3	52.3	52.5	53.2	53.6	26.2	28.5	22.3	24.6	25.4
Ours	**57.1**	**54.1**	**54.0**	**54.6**	**55.0**	**27.7**	**30.7**	**24.6**	**26.3**	**27.3**

Table 1: Performance comparison with OSOD methods on the COCO dataset in terms of AP_{50} score (%). * represents our re-implemented results with the code released by the authors.

Method	Base class																Novel class				
	plant	sofa	tv	car	bottle	boat	chair	person	bus	train	horse	bike	dog	bird	mbike	table	mAP	cow	sheep	cat	aero	mAP
CoAE	24.9	50.1	58.8	64.3	32.9	48.9	14.2	53.2	71.5	74.7	74.0	66.3	75.7	61.5	68.5	42.7	55.1	78.0	61.9	72.0	43.5	63.8
AIT	46.4	**60.5**	68.0	73.6	49.0	**65.1**	26.6	68.2	**82.6**	85.4	**82.9**	77.1	82.7	71.8	75.1	**60.0**	67.2	**85.5**	**72.8**	80.4	50.2	72.2
BHRL	**57.5**	49.4	76.8	80.4	61.2	58.4	48.1	83.3	74.3	87.3	80.1	81.0	**87.2**	73.0	78.8	38.8	69.7	81.0	67.9	**86.9**	59.3	**73.8**
BHRL*	57.0	53.1	77.9	**82.4**	61.2	59.1	48.0	**83.7**	75.0	86.9	80.8	**81.3**	85.7	70.2	76.1	41.2	70.0	81.0	65.9	85.4	58.1	72.6
Ours	55.0	55.6	**78.3**	81.4	**62.5**	59.5	**50.9**	81.7	74.7	**87.5**	82.1	81.0	85.2	**73.9**	**79.1**	39.5	**70.5**	80.6	67.4	84.6	**61.4**	73.5

Table 2: Performance comparison with OSOD methods on the PASCAL VOC dataset in terms of AP_{50} score (%). * represents our re-implemented results with the code released by the authors.

2022). As illustrated in Fig. 3, for groups (a-d), given different query images for the same target image, the corresponding target regions in the prior maps can be respectively activated, which validates the powerful generalization ability. The target images in groups (e-f) include multiple small objects or large objects, respectively. Our prior map can serve as an indicator that coarsely locates the objects of interest despite the complex scenarios containing scale variations and appearance changes, which proves the effectiveness and robustness of our module. Notably, the above steps are independent of the training process to maintain generalization ability and mitigate the bias problem.

Then, we normalize the values of the prior map to between 0 and 1, and reshape the map to match the shape of target feature $\phi(I)$ using an interpolation operation. Finally, the prior map is treated as guidance to reweight the target feature in the ROI head, facilitating the subsequent proposal features retrieval and novel-class classification.

$$\widehat{\phi}(I) = \mathcal{F}_{reshape}(\mathcal{F}_{norm}(\mathcal{P})) \odot \phi(I), \tag{18}$$

where $\odot$ stands for element-wise multiplication.

Experiments

Datasets and Settings

Datasets and Evaluation Metrics. For a fair comparison, we follow the protocol in previous works (Hsieh et al. 2019; Chen, Hsieh, and Liu 2021; Zhao, Guo, and Lu 2022; Yang et al. 2022) to implement our method. Our model is trained and tested on two widely used datasets, namely COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010). Following previous works, we report AP_{50} to evaluate performance on these two datasets.

Implementation Details. Our training process consists of two phases, base-training and novel-training. Specifically, for the base-training phase, we adopt Faster R-CNN (Ren et al. 2015) with FPN (Lin et al. 2017) as our base framework, and ResNet-50 as our backbone. In line with previous works (Hsieh et al. 2019; Yang et al. 2022), the backbone is pre-trained on the reduced ImageNet (Deng et al. 2009), where we remove all COCO-related classes to guarantee that the model does not foresee the novel-class objects. Following the general object detection paradigm, we train the base-class predictor on each group of base classes for 15 epochs. For the second phase, the parameters of the base predictor and the PG module are fixed, and the backbone weights of the novel predictor are initialized from the base predictor. The parameters of the novel predictor and the BcS module are further optimized for 8 epochs. For both phases, we adopt SGD as the optimizer with a batch size of 16, and a momentum of 0.9. The learning rate starts at 0.02 and decays by a factor of 10 at the 7th epoch.

Target-query Pairs. We apply the same strategy as (Hsieh et al. 2019; Yang et al. 2022) to generate the target-query image pairs. In the novel-training stage, given a target image from the datasets, we randomly choose one query patch containing any of the same base-class in the target image. In the testing stage, for each novel-class in a target image, the query images of the same class are shuffled with a random seed of target image ID, then the first five query images are respectively chosen to pair with the target image. We evaluate the model on these image pairs and take the average of scores as the stable results.

Comparison with State-of-the-art Methods

Comparison with OSOD Methods. Our approach is mainly oriented towards complex scenarios involving multiple categories. The challenging COCO dataset, which typically contains diverse objects, is suitable for validating our

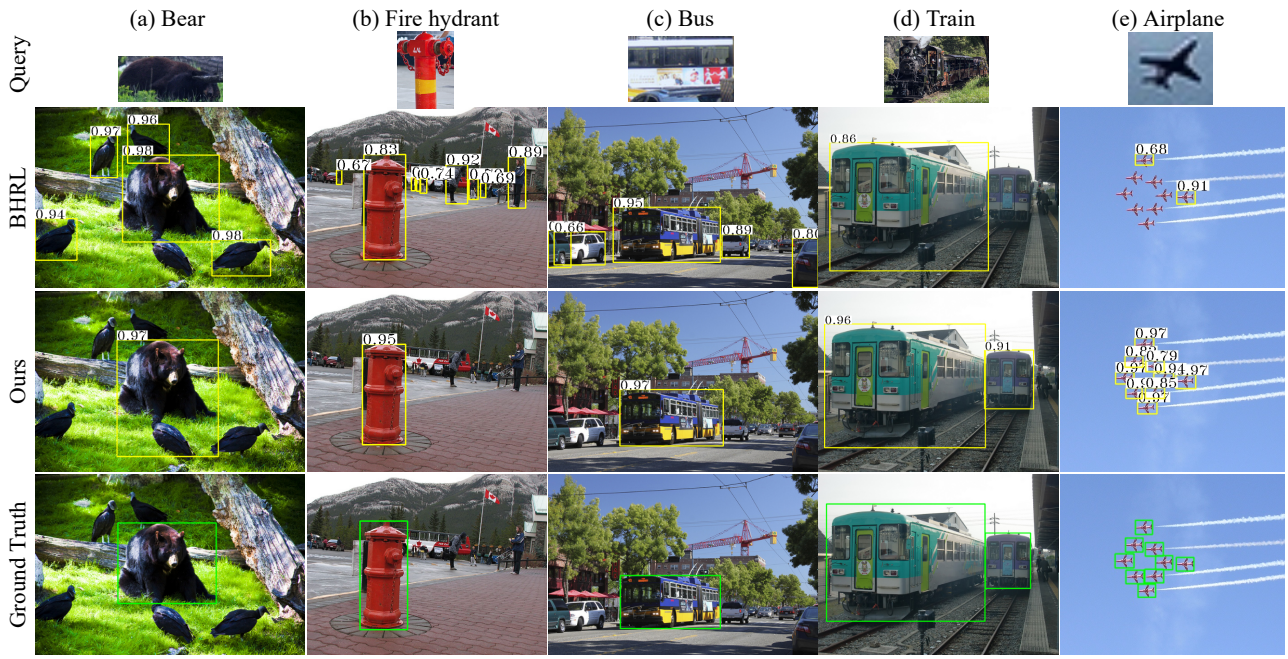

Figure 4: Comparison of qualitative one-shot object detection results for novel classes between BHRL (Yang et al. 2022) and our proposed approach. Each row from top to bottom denotes the query image, the detection results of BHRL and our model (yellow box), and ground truth (green box). Groups (a-d) are from COCO dataset, and group (e) is from PASCAL VOC dataset.

ideas. Following the common practice (Yang et al. 2022), we equally divide the 80 classes into four groups, and take three groups as base classes and one group as novel classes in turns. The results are presented in Tab. 1, our model significantly outperforms the state-of-the-art BHRL by 1.4% AP_{50} on base classes and 1.9% AP_{50} on novel classes, demonstrating its remarkable ability to handle complex scenarios. For the PASCAL VOC dataset, the 20 classes are divided into 16 base and 4 novel classes (Yang et al. 2022). As shown in Tab. 2, the proposed method achieves 0.5% and 0.9% AP_{50} improvements over the best model BHRL on the base and novel classes, respectively. We believe that the gains mainly stem from the explicit suppression of base-class distractors, which is especially effective in scenes consisting of both base and novel classes. Since most images in the PASCAL VOC dataset exhibit relatively simple scenarios with no distractors from base classes, as shown in group (e) of Fig. 4. As expected, the improvements are not as significant as that on the COCO dataset.

Qualitative Results. To better comprehend our proposed model, we visualize the detection results in Fig. 4. In groups (a-c), the baseline method is easily distracted by base-class objects and tends to misclassify the distractors as objects of interest. In contrast, our model exhibits great superiority in distinguishing target objects from distractors, which is attributed to the base-class suppression module. Besides, the crowded scenes and appearance variations are also challenges for OSOD task. As shown in groups (d-e), the baseline method neglects some objects in crowded scenes. While our model successfully identifies multiple objects and accurately localizes them, despite significant variations in appearance, scale, and shape between query and target images.

Comparison with FSOD Methods. Additionally, our model can be easily extended to few-shot settings. When several query images are available, we simply take the average of multiple query features before interacting with target features. We compare our BSPG with other advanced FSOD methods on the COCO dataset and strictly adopt the same protocol to ensure a fair comparison. Note that some FSOD algorithms (Xiao and Marlet 2020; Zhang et al. 2022a; Qiao et al. 2021) consider the task as a multi-classification and localization problem. They first train the model on abundant base data and then fine-tune it on limited novel data. The dependence on the fine-tuning process somewhat limits the application scenarios. While we treat the task as a two-classification and localization problem (Yang et al. 2022), similar to (Fan et al. 2020a; Chen et al. 2021), to better simulate the real-world application. Our approach only uses the base-class data for training and directly predicts novel-class objects guided by a query image without fine-tuning. The experiments are implemented with the same data split as in (Xiao and Marlet 2020; Zhang et al. 2022a; Qiao et al. 2021; Fan et al. 2020b,a; Chen et al. 2021), where the 20 categories overlapped with PASCAL VOC are treated as novel classes. As shown in Tab. 3, the proposed approach achieves state-of-the-art results among all settings, and surpasses the previous best competitor DAnA (Chen et al. 2021) by 3.5%, 4.6%, 5.0% AP_{50} under 1-shot, 3-shot, and 5-shot settings, respectively, validating the superiority of our approach.

The consistent gains in both OSOD and FSOD fields in-

Method	Category of classification	1-shot			3-shot			5-shot		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
FSDetView (Xiao and Marlet 2020)	Multi-classification	4.5	12.4	2.2	7.2	18.7	3.7	10.7	24.5	6.7
Meta-DETR (Zhang et al. 2022a)	Multi-classification	7.5	12.5	7.7	13.5	21.7	14.0	15.4	25.0	15.8
DeFRCN (Qiao et al. 2021)	Multi-classification	9.3	-	-	14.8	-	-	16.1	-	-
FGN† (Fan et al. 2020b)	Two-classification	8.0	17.3	6.9	10.5	22.5	8.8	10.9	24.0	9.0
Attention RPN† (Fan et al. 2020a)	Two-classification	8.7	19.8	7.0	10.1	23.0	8.2	10.6	24.4	8.3
DAnA (Chen et al. 2021)	Two-classification	11.9	25.6	10.4	14.0	28.9	12.3	14.4	30.4	13.0
Ours	Two-classification	**15.5**	**29.1**	**14.8**	**18.3**	**33.5**	**17.9**	**19.1**	**35.4**	**18.2**

Table 3: Performance comparison with FSOD methods on the COCO dataset for novel class in terms of AP, AP_{50} and AP_{75}(%). † represents the results reported in DAnA (Chen et al. 2021).

BcS	PG	AP_{50}
		28.5
✓		30.0
	✓	29.4
✓	✓	**30.7**

Table 4: Ablation study for each module in our model.

λ_1	λ_2	AP_{50}
1	0	27.8
0.5	0.5	**30.7**
0	1	29.9
1	1	30.1

Table 5: Ablation study for different values of λ_1 and λ_2 in the loss function.

Method	IoU>0.5, score>0.5		IoU>0.75, score>0.75	
	FPs	Precision (%)	FPs	Precision (%)
Before BcS	3303	41.6	725	50.0
After BcS	1894	52.6	413	66.8

Table 6: The number of FPs with respect to base classes and precision before and after BcS module, where the IoU denotes intersection over union between the final predicted results and ground truth, and the score denotes predicted score.

dicate that our model possesses a superior detection capability to identify novel objects, benefiting from the effective base-class suppression and unbiased prior guidance. Thus, resolving the model bias towards base classes is a promising direction worth exploring.

Ablation Study

We conduct a series of ablation studies on the split-2 of the COCO dataset for novel classes under 1-shot setting following the previous works (Hsieh et al. 2019; Yang et al. 2022).

Impact of Each Module. Tab. 4 shows the impact of the proposed Base-class Suppression (BcS) and Prior Guidance (PG) module on overall performance. Compared with the baseline, the BcS module brings a decent performance gain of 1.5% AP_{50}, demonstrating its effectiveness in suppressing the false detections on the base-class objects. The comparison of rows 1 and 3 shows that the PG module contributes to a 0.9% AP_{50} improvement over the baseline. By combining all the modules, we achieve the best results and exceed the baseline by 2.2% AP_{50}, providing convincing proof that our proposed modules indeed enhance the ability to detect novel objects. Specifically, the BcS module can effectively suppress the distractors to explicitly resolve the model bias towards base classes, and the non-parametric PG module can generate unbiased prior knowledge to implicitly mitigate the bias problem.

Impact of Parameter λ_1 and λ_2 in the Loss Function. In the loss function defined by Eq. (15), λ_1 and λ_2 are the weights assigned to $\mathcal{L}_N^{RP}$ and $\mathcal{L}_F^{RP}$, respectively. We investigate the impact of λ_1 and λ_2 on the final performance in Tab. 5. The comparison of rows 1 and 4 reveals the indispensable role of supervision on the final results. Moreover, imposing supervision on the intermediate results predicted by the novel predictor can facilitate further refinement and boost the final performance. Based on our experiments, we can conclude that when λ_1 and λ_2 are set to 0.5, our model yields the best performance.

Impact of BcS Module on the False Prediction. The BcS module is a core component of our model, serving to eliminate false positives (FPs) and explicitly address the bias problem. To further analyze its impact, we report the number of FPs with respect to base classes both before and after applying the BcS module in Tab. 6. Precision, formulated as $\frac{TP}{TP+FP}$, is also used to assess the performance concerning the suppression of FPs. Note that the experiments are conducted on the split-2 of the COCO dataset, which contains 3309 test images. The results clearly indicate that the number of FPs is substantially reduced, and the precision is remarkably improved after the BcS module, thus verifying the effectiveness of the BcS module.

Conclusion

In this paper, targeting the phenomenon of model bias towards base classes and the generalization degradation on the novel classes, we rethink one-shot object detection from a new perspective and propose a BSPG network to achieve bias-free OSOD. We design a base-class predictor and a base-class suppression module to explicitly recognize and suppress base-class objects. Furthermore, the PG module is proposed to generate a class-agnostic prior map with unbiased semantic hints to guide the detection procedure and enhance overall performance. Extensive experiments demonstrate that our approach reaches new state-of-the-art performance under all settings. We believe that our work offers valuable insights into alleviating the bias problem in the OSOD field and can inspire future research in this area.

Acknowledgments

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LGF20F010006, and in part by the National Natural Science Foundation of China under Grant U21B2029 and Grant U21A20456.

References

Chen, D.-J.; Hsieh, H.-Y.; and Liu, T.-L. 2021. Adaptive Image Transformer for One-Shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12247–12256.

Chen, T.-I.; Liu, Y.-C.; Su, H.-T.; Chang, Y.-C.; Lin, Y.-H.; Yeh, J.-F.; Chen, W.-C.; and Hsu, W. 2021. Dual-Awareness Attention for Few-Shot Object Detection. *IEEE Transactions on Multimedia*, 1–1.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. CenterNet: Keypoint Triplets for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020a. Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, Z.; Ma, Y.; Li, Z.; and Sun, J. 2021. Generalized Few-Shot Object Detection Without Forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4527–4536.

Fan, Z.; Yu, J.-G.; Liang, Z.; Ou, J.; Gao, C.; Xia, G.-S.; and Li, Y. 2020b. FGN: Fully Guided Network for Few-Shot Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Han, G.; Huang, S.; Ma, J.; He, Y.; and Chang, S.-F. 2022. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 780–789.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019. One-Shot Object Detection with Co-Attention and Co-Excitation. In *Advances in Neural Information Processing Systems*, volume 32.

Lang, C.; Cheng, G.; Tu, B.; and Han, J. 2022. Learning What Not To Segment: A New Perspective on Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8057–8067.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Michaelis, C.; Ustyuzhaninov, I.; Bethge, M.; and Ecker, A. S. 2018. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8681–8690.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7352–7362.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; and Jia, J. 2022. Prior Guided Feature Enrichment Network for Few-Shot Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 1050–1065.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29: 3630–3638.

Xiao, Y.; and Marlet, R. 2020. Few-shot object detection and viewpoint estimation for objects in the wild. In *European conference on computer vision*, 192–210. Springer.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yang, H.; Cai, S.; Sheng, H.; Deng, B.; Huang, J.; Hua, X.-S.; Tang, Y.; and Zhang, Y. 2022. Balanced and hierarchical relation learning for one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7591–7600.

Yang, H.; Lin, Y.; Zhang, H.; Zhang, Y.; and Xu, B. 2021. Towards improving classification power for one-shot object detection. *Neurocomputing*, 455: 390–400.

Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; and Xing, E. P. 2022a. Meta-DETR: Image-Level Few-Shot Detection with Inter-Class Correlation Exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12.

Zhang, W.; Dong, C.; Zhang, J.; Shan, H.; and Liu, E. 2022b. Adaptive context- and scale-aware aggregation with feature alignment for one-shot object detection. *Neurocomputing*, 514: 216–230.

Zhao, Y.; Guo, X.; and Lu, Y. 2022. Semantic-Aligned Fusion Transformer for One-Shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7601–7611.