FOC OSOD: FOCUS ON CLASSIFICATION ONE-SHOT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-shot object detection (OSOD) aims at detecting all instances that are consistent with the category of the single reference image. OSOD achieves object detection by comparing the query image and the reference image. We observe that the essential problem behind the limited performance of OSOD is that OSOD generates a lot of false positives due to its poor classification ability. This paper analyzes the serious false positive problem in OSOD and proposes a Focus on Classification One-Shot Object Detection (FOC OSOD) framework, which is improved in two important aspects: (1) classification cascade head with the fixed IoU threshold can enhance the robustness of classification by comparing multiple close regions; (2) classification region deformation on the query feature and the reference feature to obtain a more effective comparison region. Without bells and whistles, a single FOC obtains 1.8% AP and 1.3% AP improvement on the seen classes and the unseen classes over a Siamese Faster R-CNN baseline on the MS-COCO dataset in the one-shot setting. The code will be available.

1 INTRODUCTION

In recent years, object detection methods based on convolutional neural networks(25; 2) have achieved great success. However, this success relies on a large training dataset with laborious labelings, such as MS-COCO(19), which can only detect the categories annotated in the training set. This makes the general object detection methods difficult to extend new object categories. For one thing, it is time-consuming to perform the much annotation work and errors often exist during the labeling; for another, it is difficult to collect a large number of images of new categories in some special scenes. Therefore, it is valuable and necessary to enable the model to detect the unseen category when only a small number of images are provided.

The previous works(21; 17) use siamese structure(5; 15) to this task, where SiamMask(21) attains this by adding a matching layer to Mask R-CNN(10), CoAE(17) uses the non-local(29) and co-excitation(13) to enhance the correlation between the reference feature and the query feature. To the best of our knowledge, the false positive problem in one-shot obeject detection (OSOD) has not been studied.

In this paper, through the preliminary experiments, we find that when the classification branch does not introduce the reference feature information, it can cause more obvious performance degradation due to more false positives detected than the regression branch without the reference feature information. The above observation stimulates our work to improve the classification power of the OSOD. A Focus on Classification One-Shot Object Detection (FOC OSOD) framework is presented in this paper, where a classification cascade head is designed to enhance the robustness of classification by comparing multiple close regions. Instead of training the detector of the next stage higher IoU threshold(2), we use the fixed IoU threshold in different stages. Moreover, the detector needs to know which candidate regions in the query image and the reference image are more effective to be compared. We propose the classification region deformation on the query feature and the reference feature that obtains the more effective comparison region. Experiments show that a single FOC achieves an increase in the seen classes and the unseen classes, 1.8% AP and 1.3% AP respectively, over a Siamese Faster R-CNN baseline on MS-COCO dataset in the one-shot setting.

The main contributions of this work can be summarized as follows:

- As far as we know, it is the first work to discuss the false positive problem in OSOD.
- To solve the problem, we propose a FOC OSOD framework. To be specific, we present the classification cascade head with the fixed IoU threshold, and the classification region deformation on the query feature and the reference feature to improve the classification power.
- Extensive experiments demonstrate that our method outperforms the previous state-of-the-art counterparts on PASCAL VOC(8) and MS-COCO datasets. Moreover, ours improves by 1.8% AP and 1.3% AP on the seen classes and the unseen classes over a Siamese Faster R-CNN baseline on the MS-COCO dataset.

2 Related Work

2.1 Few-Shot Object Detection

Few-shot object detection aims to recognize novel objects given several or even one reference image, which is challenging. In recent years, some work has focused on this task. LSTD(3) builds a detector that fine-tunes on the target domain by transferring knowledge and background depression regularizations from the source and target datasets. RepMet(26) applies the distance metric learning into the RoI classification head in the detector to map the objects to the embedding space. All the above-mentioned methods need to fine-tune in the target datasets. CompNet(32) compares the reference and query features with the learnable metric to optimize the non-linear conditional probability. CoAE(12) uses the non-local operation(29) and the squeeze-and-co-excitation scheme(13) to explore the correlated feature in the reference and query. FSOD(9) introduces depth-wise convolution to get the attention feature map in the RPN phase and proposes the multi-relation detector to model different relationships in the R-CNN phase. (17) adopts a first stage Matching-FCOS network to increase the recall and a second stage structure-aware relation module to improve the precision. OS2D(22) presents dense correlation matching of local features and performs spatial alight and bilinear resampling to compute the detection score. (30) highlights the importance of scale variations and generates multi-scale samples to enrich object scales. However, these methods are only considering how to highlight the correlation between the reference and the query image. To the best of our knowledge, it is the first work to discuss the false positive problem in OSOD.

2.2 General Object Detection

Object detection is one of the basic tasks in computer vision, which has seen remarkable progress in recent years. Unlike few-shot object detection, general object detection needs to localize all objects from a series of fixed categories. The methods can be divided into two types of pipeline designs: two-stage(25; 2) and one-stage(23; 24; 20; 18; 28). Most two-stage methods follow Faster R-CNN(25), which first generates proposals by region proposal network (RPN) and then performs detection within each proposal. In contrast, the one-stage methods like SSD(20), YOLO(23), RetinaNet(18), FCOS(28) have a faster inference speed but often less accurate than two-stage methods. Recently, a few works(4; 14; 31; 27) are conscious about the conflict between the classification and localization, which limits the performance of the network. IoU-Net(14) adds an extra head to predict the IoU and then combines the classification and the localization confidence as the final classification score, which improves the NMS procedure by preserving accurately localized bounding boxes. Double-Head R-CNN(31) decouples the sibling head into two special heads for classification and localization, respectively. They found that the convolutional head is more suitable for the regression task, and the fully connected head is more proper for the classification task. (4) designs a decoupled classification refinement (DCR) module to improve the classification power and eliminate high-scored false positives. TSD(27) solves the problem by decoupling the classification and regression to generate two different proposals. However, there is no previous work that has mentioned the focus of the reference image on classification and regression for OSOD.

3 Our Method

In this section, we discuss our FOC OSOD approach in details. Firstly, we introduce our hypothesis about the focus of the reference feature on the classification task and the regression task for OSOD

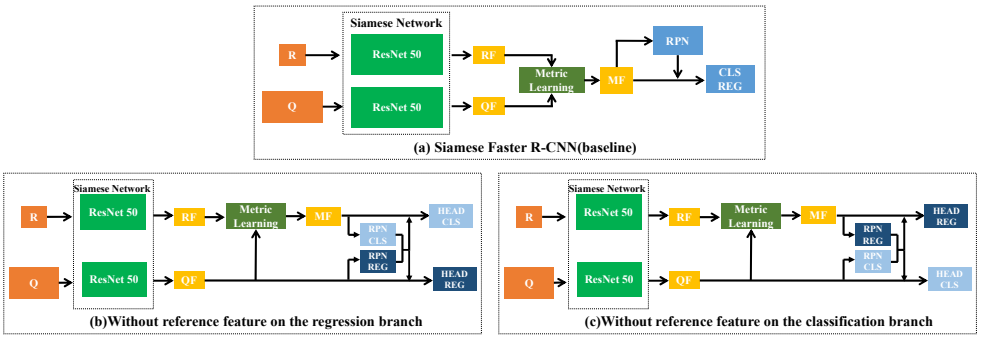

Figure 1: The architecture of different networks for preliminary experiments. "R", "Q", "RF", "QF", "MF" denote reference image, query image, reference feature, query feature, metric-learning feature, respectively.

and provide a deep analysis of the hypothesis in the preliminary experiments. Then, the complete structure of the proposed FOC OSOD is introduced. Next, the classification cascade head with the fixed IoU threshold is presented, which can enhance the robustness of classification by comparing multiple close regions. Finally, we give the classification region deformation on the query feature and the reference feature, which can get the more effective comparison region.

3.1 Motivation

Unlike general object detection, OSOD needs to detect objects of the unseen categories with only one annotated image. The main difference between OSOD and general object detection is that we need to mine the information of the reference image in OSOD. As it is observed, lots of false positives are detected in a cluttered background. In other words, the network can locate the objects from the background, but it cannot accurately separate the objects with the same category as the reference image from all the objects. This motivates us to think about the focus of the reference feature on the classification task and the localization task for OSOD. Given the motivation discussed above, we have a hypothesis as follows: the reference feature is more important to the classification branch than the regression branch for OSOD.

To validate this hypothesis, we adopt the Siamese Faster R-CNN as our baseline for the preliminary experiments. Figure 1 (a) shows the Siamese Faster R-CNN network, which is built by removing the semantic branch of SiamMask(21). Figure 1 (b) and Figure 1 (c) show the structure of the network without the reference feature on the regression branch or the classification branch, respectively, where we decouple the classification branch and the regression branch for the RPN and head to eliminate the information sharing of reference features in RPN and head. Figure 2(a) shows the comparison of the number of false positives in different confidence scores ranges and Figure 2(b) illustrates the mAP comparison across IoU thresholds from 0.5 to 0.8. The following observations on the experiments are made: (1) The OSOD model with fewer false positives has higher accuracy. (2) The reference feature plays a more important role in the classification task rather than the localization task. (3) Our method can improve accuracy by reducing the number of false positives.

3.2 Network Architecture

As shown in Figure 3, we build FOC OSOD based on Siamese Faster R-CNN. To be more specific, given a query image and a reference image, we extract features by a shared weight siamese network and exploit the similarity between the query image and the reference image by a metric learning layer(21). The region proposal network (RPN) is used to produce potentially relevant boxes to facilitate the following task of the detection. In the heads, we recalculate the similarity map for each proposed region, to be more specific, we utilize RoI-Align(10) on the reference feature and the query selected feature to obtain the pooling features with the same shape, and then propose the classification region deformation on the two pooling features to get a more effective classification comparison region. Next, the metric learning layer is used on the new features to get a more effective

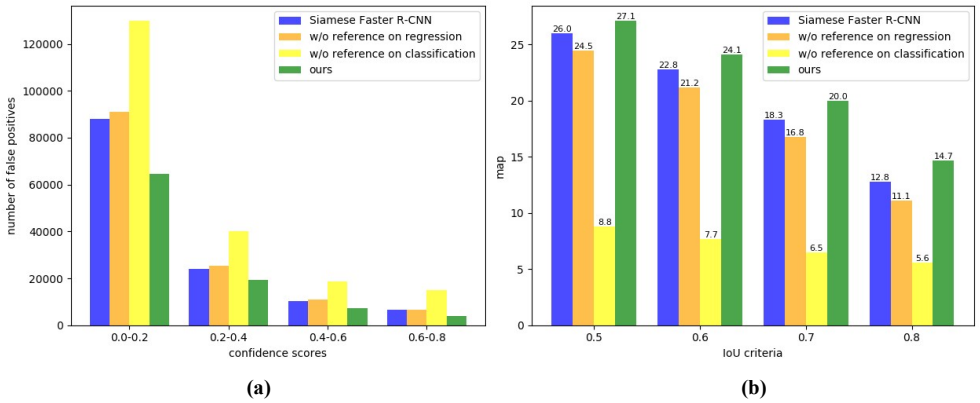

Figure 2: Illustration for the poor performance of OSOD due to a large number of false positives.

similarity map. Then, to reduce the influence of the regression branch on the classification branch, we separate the classification and regression branch, following (31), we apply the fully connected head on the classification task and the convolution head on the regression task. Instead of using four residual bottleneck blocks on the regression task, only one residual bottleneck block is used for saving memory. Finally, we propose the classification cascade head with the fixed IoU threshold to improve the classification power.

3.3 Classification Cascade Head with the Fixed IoU Threshold

The OSOD setting can result in lots of false positives because the only reference feature is difficult to represent the information of all instances with the same category. To resolve this matter, we improve the classification power by a classification cascade operation. Firstly, like the baseline, we apply the pooling operator on the metric-learning features, which is formulated as:

$$\begin{aligned} \mathrm{MF_r} &= \mathcal{P}(\mathrm{MF}, \mathrm{p_r}), \\ \mathrm{MF_h} &= \mathcal{P}(\mathrm{MF}, \mathrm{p_h}). \end{aligned} \tag{1}$$

Here, MF indicates the metric-learning features generated before RPN. $\mathrm{p_r}$ and $\mathrm{p_h}$ represent the box predictions generated by the RPN and the head, respectively. $\mathcal{P}(\cdot, \cdot)$ is the RoI-Align operator. $\mathrm{MF_r}$ and $\mathrm{MF_h}$ denote the pooling features which is input to the first and second head stage, respectively.

Then the $\mathrm{MF_r}$ is input to the first stage classification head and regression head, and the $\mathrm{MF_h}$ is input to the second stage classification head. We use the same IoU threshold (e.g. 0.5) to train the classification cascade head for novel class learning because we argue that the detector of the next stage higher IoU threshold is suitable for box optimization instead of the classification optimization. The $\mathrm{p_r}$ and $\mathrm{p_h}$ are similar which are both close to ground truth, but not the same. In this way, we can compare whether the query and reference are the same category from multiple close regions, and then weigh the outputs to improve the robustness. Compared with the cascade method proposed in (2), the main difference is that our method does not aim to optimize the appropriate quality proposal bounding box by training the detector of the next stage higher IoU threshold. Instead, the classification cascade head with the fixed IoU threshold aims to enhance the robustness of classification by comparing multiple close regions. The head execution is expressed as:

$$\begin{aligned} \mathrm{s_1} &= \mathcal{F}_{cls1}(\mathrm{MF_r}), \\ \mathrm{s_2} &= \mathcal{F}_{cls2}(\mathrm{MF_h}), \\ \mathrm{p_h} &= \mathcal{F}_{reg}(\mathrm{MF_{\hat{r}}}). \end{aligned} \tag{2}$$

where $\mathcal{F}_{cls1}$ and $\mathcal{F}_{cls2}$ are the first and second stage classification functions, which is the three-layer fully connected network with output $\{1024,1024,2\}$ for each layer. $\mathcal{F}_{reg}$ is the regression function using the convolution operator, which is comprised of one residual block and one residual bottleneck block (31). $\mathrm{s_1}$ and $\mathrm{s_2}$ indicate the first and second stage classification score.

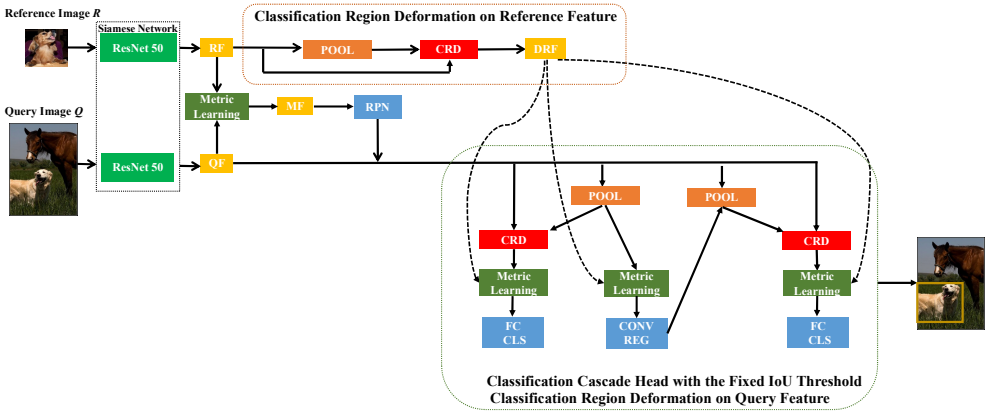

Figure 3: The overall network architecture of the proposed FOC OSOD.

3.4 Classification Region Deformation

In the design above, the classification prediction at each stage is based purely on the MF generated before RPN. There is no direct instance comparison between the reference feature and the query feature, which prevents further improvements in classification accuracy. Towards a good design of classification comparison, we first crop the query feature and the reference feature into NxN fixed-size feature maps, and then the query feature map and the reference feature map are processed by the metric-learning. Such an instance comparison structure is more suitable to compare and classify. Moreover, it is critical to select the effective regions in the query and reference features to compare whether they belong to the same class. The issue arises to think about whether the pooling query and reference feature maps by RoI Align can generate high-quality features for classification. We embed the deformation-learning manner(6) to perform classification region deformation for both the reference and query feature maps. With the classification region deformation on the query feature and the reference feature, the execution can be written as:

$$\begin{aligned} \mathrm{QF_r} &= \mathcal{F}(\mathcal{P}(\mathrm{QF}, \mathrm{p_r}), \mathrm{QF}, \mathrm{p_r}), \\ \mathrm{QF_h} &= \mathcal{F}(\mathcal{P}(\mathrm{QF}, \mathrm{p_h}), \mathrm{QF}, \mathrm{p_h}), \\ \mathrm{RF_d} &= \mathcal{F}(\mathcal{P}(\mathrm{RF}, \mathrm{r}), \mathrm{RF}, \mathrm{r}). \end{aligned} \tag{3}$$

where QF denotes the query feature of the backbone network and the RF denotes the reference feature of the backbone network. r indicates the region of the reference feature without padding. $\mathcal{F}$ is the function to perform the classification region deformation.

Following the discussion above, the $\mathcal{F}$ implementation is as below:

$$\mathcal{F}(\mathcal{P}(\mathrm{QF}, \mathrm{p_r}), \mathrm{QF}, \mathrm{p_r}) = \mathcal{F}_{\mathrm{for}}\Big(\sum_{\mathrm{p} \in \mathrm{G(x,y)}} \frac{\mathcal{BI}((\gamma \mathcal{F}_c(\mathcal{P}(\mathrm{QF}, \mathrm{p_r})) \cdot (\mathrm{w}, \mathrm{h}) + \mathrm{p}), \mathrm{QF})}{\mathrm{n_{xy}}} \Big) \tag{4}$$

where $\mathcal{F}_{\mathrm{for}}$ denotes the loop for each grid in the p_r, G(x,y) indicates the (x,y)-th grid and n_{xy} indicates the number of pixels in the grid. $\mathcal{F}_c$ is the function to obtain the offsets based on the pooling feature maps, which is a three-layer fully connected network with output {256, 256, NxNx2} for each layer. γ is the pre-defined scalar to modulate the magnitude of the offsets and the (w,h) is the width and height of p_r. $\mathcal{BI}$ denotes the bilinear interpolation.

Then, the metric-leaning process can be formulated as:

$$\begin{aligned} \mathrm{MF_r} &= \mathrm{Conv}_{1,384}(\mathrm{QF_r} \odot (\mathrm{QF_r} - \mathrm{GAP}(\mathrm{RF_d}))) \in \mathbb{R}^{384 \times \mathrm{N} \times \mathrm{N}}, \\ \mathrm{MF_{\hat{r}}} &= \mathrm{Conv}_{1,384}(\mathcal{P}(\mathrm{QF}, \mathrm{p_r}) \odot (\mathcal{P}(\mathrm{QF}, \mathrm{p_r}) - \mathrm{GAP}(\mathrm{RF_d}))) \in \mathbb{R}^{384 \times \mathrm{N} \times \mathrm{N}}, \\ \mathrm{MF_h} &= \mathrm{Conv}_{1,384}(\mathrm{QF_h} \odot (\mathrm{QF_h} - \mathrm{GAP}(\mathrm{RF_d}))) \in \mathbb{R}^{384 \times \mathrm{N} \times \mathrm{N}}. \end{aligned} \tag{5}$$

Here, $\odot$ refers to the concatenation operation and the GAP is the global average pooling. We denote a convolution layer with kernel size s as $\mathrm{Conv}_{s,n}(\cdot)$, where *n* is the output number of kernels.

Finally, the classification and regression are implemented via Eq. 2. The overall loss function takes the form of multi-task learning:

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{reg} + \mathcal{L}_{cls1} + \mathcal{L}_{cls2} \tag{6}$$

4 EXPERIMENTS

In this section, the experimental results are presented to evaluate the effectiveness of the proposed methods on PASCAL VOC(8) and MS-COCO(19) datasets. For fair quantitative comparison with the state-of-the-art method, we follow the setups in (17) to construct the one-shot detection datasets. All of our ablation studies are conducted on MS-COCO datasets.

4.1 IMPLEMENTATION DETAILS

In the experiments, by default we train our model according to the following settings unless otherwise stated. We use stochastic gradient descent (SGD) over eight NVIDIA RTX 2080 Ti GPUs with a total of 16 images per mini-batch. Our model is trained for 10 epochs with an initial learning rate of 0.02, which is then divided by 10 at 7th and again at 10th epoch. Our network is trained and tested with the 1024x1024 query image and the 192x192 reference image. We use the pre-trained model ResNet-50(11) from (17). The ResNet-50 model is trained on a reduced training set of ImageNet(7) which removes all classes that are related to COCO, which is to ensure that model does not 'foresee' the unseen classes. We run all our evaluations five times and average the results for stability as (17). To save the training time, except for the comparison experiments with the state-of-the-art methods to be trained on all four COCO splits, the other ablation experiments are carried out on COCO split2.

4.2 MAIN RESULTS

We train and evaluate our model on PASCAL VOC and MS-COCO benchmark datasets.

Comparison on PASCAL VOC Following the dataset setting in the previous work (17), the 20 classes in PASCAL VOC datasets are split into 16 seen classes and 4 unseen classes. Our model is trained on the union set of VOC 2007 train&val sets and VOC 2012 train&val sets, and is evaluated on VOC 2007 test set. In the training, we only train the seen classes, while in the test, we test the seen classes and the unseen classes respectively, and calculate the average precision (AP) of each category. The experimental results are summarized in Table 1. It can be seen from this table that our network outperforms the other methods by a large margin on both seen and unseen classes. It shows that our network outperforms CoAE by 11.1% AP on seen classes and by 5.3% AP on unseen classes. Furthermore, the better performance on the unseen classes than the seen classes shows that our model can easily detect novel unseen instances.

Table 1: Comparison of diffenrent methods on PASCAL VOC in detection AP_{50}.

Model	Seen class																	Unseen class				
	plant	sofa	tv	car	bottle	boat	chair	person	bus	train	horse	bike	dog	bird	mbike	table	mAP	cow	sheep	cat	aero	mAP
SiamFC(1)	3.2	22.8	5.0	16.7	0.5	8.1	1.2	4.2	22.2	22.6	35.4	14.2	25.8	11.7	19.7	27.8	15.1	6.8	2.28	31.6	12.4	13.3
SiamRPN (16)	1.9	15.7	4.5	12.8	1.0	1.1	6.1	8.7	7.9	6.9	17.4	17.8	20.5	7.2	18.5	5.1	9.6	15.9	15.7	21.7	3.5	14.2
CompNet(32)	28.4	41.5	65.0	66.4	37.1	49.8	16.2	31.7	69.7	73.1	75.6	71.6	61.4	52.3	63.4	39.8	52.7	75.3	60.0	47.9	25.3	52.1
CoAE(12)	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
Ours	**55.7**	45.1	**74.4**	**78.2**	**57.2**	**56.9**	**46.3**	**81.7**	67.3	**83.3**	**79.7**	**80.9**	**83.0**	**70.5**	**76.0**	23.5	**66.2**	**78.0**	**62.6**	**82.8**	**52.9**	**69.1**

Comparison on MS-COCO We show the results on the challenging MS-COCO benchmark, which contains 118k training images and 5k validation images. We adopt the same data setting as (17), the 80 classes in COCO datasets are split into 60 train categories and 20 test categories. Four such training/test splits are generated by including every fourth category into the test category starting with the first, second, third, or fourth category (21), respectively. Following(17), we filter out the too small or too hard image patches. Table 2 and Table 3 show that the comparison with the baseline (Siamese Faster R-CNN) and CoAE. It is worth mentioning that due to better implementation and training strategy, our baseline model achieves 10.4% AP and 12.0% AP_{50} higher than

Table 2: Detection results on COCO 2017 val of seen classes.

Method	Split1		Split2		Split3		Split4		Average	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
Baseline	33.7	55.5	31.0	51.4	32.0	51.8	31.9	52.5	32.2	52.8
CoAE	22.4	42.2	21.3	40.1	21.6	39.8	22.0	41.0	21.8	40.8
Ours	**35.5**	**56.2**	**32.8**	**51.8**	**33.8**	**52.6**	**34.0**	**53.4**	**34.0**	**53.5**

Table 3: Detection results on COCO 2017 val of unseen classes.

Method	Split1		Split2		Split3		Split4		Average	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
Baseline	13.5	23.8	15.4	26.0	11.0	20.7	12.8	23.0	13.2	23.4
CoAE	11.8	23.2	12.2	23.7	9.3	20.3	9.4	20.4	10.7	21.9
Ours	**14.6**	**24.8**	**16.9**	**27.1**	**12.2**	**21.1**	**14.1**	**23.8**	**14.5**	**24.2**

CoAE on the seen classes, and achieves 2.5% AP and 1.5% AP_{50} higher than CoAE on the unseen classes. The performance of the CoAE is obtained by running their open code on the same data setting as ours. Moreover, compared with the baseline model, our FOC OSOD improves by 1.8% AP and 0.7% AP_{50} on the seen classes, and 1.3% AP and 0.8% AP_{50} on the unseen classes. Figure 4 shows the visualization comparison.

4.3 Ablation Experiments

We investigate the effects of the main components in our framework. "double head" denotes that the fully connected head is applied on the classification task and the convolution head is applied on the regression task. "ccd" denotes the classification cascade head with the fixed IoU threshold. "crd" denotes the classification region deformation on the query feature and the reference feature. From Table 4, we can learn that the double head slightly improves by 0.3% AP on the unseen classes, but drops by 1.0% on the seen classes. The classification cascade head with the fixed IoU threshold contributes to a further 2.1% and 0.6% improvement on the seen and the unseen classes, respectively. The classification region deformation leads to a gain of 0.7% and 0.6% on the seen and the unseen classes, respectively.

4.4 Effectiveness of Classification Cascade Head

We design the classification cascade head to benefit the classification branch from comparing multiple close regions during training. From Table 5, we find that introducing the classification cascade head with the fixed IoU threshold improves 2.1% AP on the seen classes and 0.6% AP on the unseen classes. We argure that using the next stage higher IoU threshold setting as (2) is not suitable for the only classification cascade head on unseen classes as the improvement is limited (0.4%). However, the fixed IoU threshold training can obtain more significant gain (0.6%). The ensemble between the first and second stage contributes to the performance.

Table 4: Ablation study on the major components on COCO 2017 val split2. The ccd denotes cascade classification head, the crd denotes classification region deformation.

				seen	unseen
baseline	double head	ccd	crd	AP	AP
✓				31.0	15.4
✓	✓			30.0	15.7
✓	✓	✓		32.1	16.3
✓	✓	✓	✓	**32.8**	**16.9**

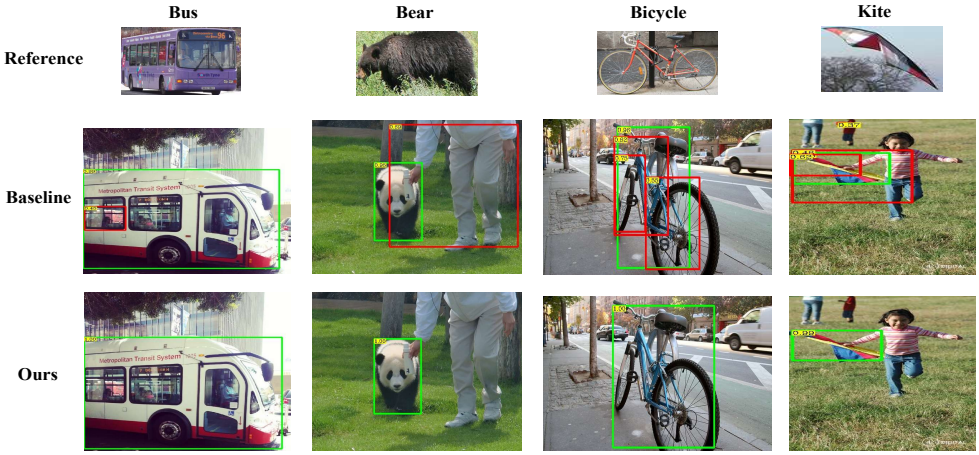

Figure 4: Comparison of false positives between baseline and ours. Green boxes indicate correctly detected objects and red boxes indicate wrong detections.

Table 5: Ablation study of cascade classification head on COCO 2017 val split2.

			seen	unseen
stage1 IoU	stage2 IoU	test stage	AP	AP
0.5	w/o	1	30.0	15.7
0.5	0.6	1-2	32.4	16.1
0.5	0.5	1	31.8	15.6
0.5	0.5	2	31.2	16.2
0.5	0.5	1-2	32.1	**16.3**

4.5 Effectiveness of Classification Region Deformation

We investigate the contributions of the classification region deformation on the query feature or the reference feature and summarize the results in Table 6. Adding the classification region deformation on the query feature boosts the performance with an increase of 0.6%/0.2% AP and 0.8%/0.6% AP_{50}on the seen and the unseen classes, respectively. Moreover, applying the classification region deformation on the reference feature has little effect on the seen classes, but it can achieve an extra 0.4% AP and 0.3% AP_{50} gain on the unseen classes. This implies that the effective comparison area between the query feature and the reference feature is helpful for OSOD.

Table 6: Ablation study of classification region deformation on COCO 2017 val split2.

CRD		seen		unseen	
query	reference	AP	AP_{50}	AP	AP_{50}
		32.1	51.0	16.3	26.2
✓		32.7	51.8	16.5	26.8
✓	✓	**32.8**	**51.8**	**16.9**	**27.1**

5 Conclusions

This paper aims to solve the false positives problem due to the poor classification power in OSOD. To deal with this issue, we propose an one-shot object detection framework to focus the reference feature on the classification task, named FOC OSOD. In particular, we design a classification cascade head with the fixed IoU threshold to improve the classification power by comparing multiple close regions and apply the classification region deformation on the query feature and the reference feature to obtain a suitable comparison region. Experiments are extensively carried out on PASCAL VOC and MS-COCO, which have shown that FOC OSOD achieves state-of-the-art results.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, A. Vedaldi, and P. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV Workshops*, 2016.

[2] Zhaowei Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[3] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. *ArXiv*, abs/1803.01529, 2018.

[4] Bowen Cheng, Yunchao Wei, Humphrey Shi, R. Feris, Jinjun Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. *ArXiv*, abs/1803.06799, 2018.

[5] S. Chopra, Raia Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1, 2005.

[6] Jifeng Dai, Haozhi Qi, Y. Xiong, Y. Li, Guodong Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[7] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[8] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009.

[9] Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. *ArXiv*, abs/1908.01998, 2019.

[10] Kaiming He, Georgia Gkioxari, P. Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[11] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *ArXiv*, abs/1911.12529, 2019.

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2020.

[14] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. *ArXiv*, abs/1807.11590, 2018.

[15] Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

[16] B. Li, J. Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.

[17] Xiang Li, Lin Zhang, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. One-shot object detection without fine-tuning. *ArXiv*, abs/2005.03819, 2020.

[18] Tsung-Yi Lin, Priyal Goyal, Ross B. Girshick, Kaiming He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.

[19] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[20] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, S. Reed, Cheng-Yang Fu, and A. Berg. Ssd: Single shot multibox detector. *ArXiv*, abs/1512.02325, 2016.

[21] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. *ArXiv*, abs/1811.11507, 2018.

[22] Anton Osokin, Denis Sumin, and Vasily Lomakin. Os2d: One-stage one-shot object detection by matching anchor features. *ArXiv*, abs/2003.06800, 2020.

[23] Joseph Redmon, S. Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[24] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[26] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Sharath Pankanti, Rogério Schmidt Feris, Abhishek Kumar, Raja Giryes, and Alexander M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5192–5201, 2019.

[27] Guanglu Song, Y. W. Liu, and Xiao gang Wang. Revisiting the sibling head in object detector. *ArXiv*, abs/2003.07540, 2020.

[28] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, 2019.

[29] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[30] Jianping Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. *ArXiv*, abs/2007.09384, 2020.

[31] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization in r-cnn. *ArXiv*, abs/1904.06493, 2019.

[32] Tengfei Zhang, Yue Zhang, Xian Sun, Hao Sun, Menglong Yan, Xue Yang, and Kun Fu. Comparison network for one-shot conditional object detection. *ArXiv*, abs/1904.02317, 2019.