Semi-Supervised Few-Shot Object Detection with A Teacher-Student Network

Wuti Xiong[1]
wuti.xiong@oulu.fi
Yawen Cui[1]
yawen.cui@oulu.fi
Li Liu *, 2, 1
li.liu@oulu.fi

[1] Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland

[2] College of System Engineering
National University of Defense Technology, China

Abstract

Few-shot object detection, which aims to recognize unseen objects with a few annotated instances, has attracted increasing attention in the computer vision community. Most recent works tackle this problem under the meta-learning framework based on an episodic training strategy. In this work, we advance the few-shot object detection paradigm towards a new scenario called semi-supervised few-shot object detection (SSFSOD), where the unlabeled data are available within each episode. To address this paradigm, we propose a novel method which utilizes a dual model (teacher-student) to leverage available unlabeled data. Specifically, the teacher model provides high-quality pseudo-labels for the student model during the training process, while the student model uses the exponential moving average strategy to update the teacher model online. We also employ a two-fold correlation-guided attention module to guide RPN to generate task-specific region proposals by highlighting potential regions and informative channels. We conduct extensive experiments on three datasets MS COCO, PASCAL VOC, and FSOD. The experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

Due to the latest boom in deep learning techniques, generic object detection [4, 9, 23, 24, 25] has made immense progress in the past decade. Recent progress in generic object detection, while substantial, has been so far limited to the form of fully supervised learning, relies on a great amount of accurately labeled images, and the performance substantially degrades when labeled data is scarce. This poses great limitations for real-world applications. Firstly, massive amounts of object bounding box annotations or segmentation masks for training detectors are expensive and time consuming to obtain. Secondly, supervised detectors have limited generalization performance due to the constrained domains defined largely by the training images. Thirdly, visual objects naturally exhibit an imbalance in their category distribution, where many minority categories such as endangered species and critical security scenarios only contain several or a few samples [28, 38]. Finally, most existing data-hungry detectors

* indicates the corresponding author.
© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

learning to detect objects are different from human learning to recognize objects. Humans can learn novel concepts from one or several examples by utilizing previously learned knowledge [42]. Therefore, to address some of these challenges, the objective of this paper is to study the problem of Few Shot Object Detection and develop object detectors that are label efficient yet are capable of rapidly generalizing to new tasks with very limited supervised information.

Recently, inspired by the success of few-shot image classification [3, 12, 17, 33, 34, 49], some studies start shifting towards a few-shot object detection problem. One family of work [11, 45] explores meta-learning, which aims to learn across tasks and then adapt to new tasks. These methods firstly learn a series of detection tasks, each of which consists of a few annotated samples in the base classes, and then adapt to the novel detection task where each class only has several annotated instances. In generic object detection, fortunately, a huge number of unlabeled images (*e.g.*, the massive amounts of unlabeled images available from the Internet) are significantly cheap to obtain. How to make good use of such unlabeled data becomes crucial. This is called Semi-Supervised Learning (SSL), which has also been extensively studied in image classification problems [21, 31, 48]. As one of the current bottlenecks in object detection is obtaining labels, in this work, our main focus is to conduct a pioneering work by developing a framework for *SemiSupervised* FSOD (SSFSOD), *i.e.*, leveraging unlabeled images for FSOD.

To address this new problem setting, *i.e., SSFSOD*, we build upon recent works on meta-learning based FSOD [11, 45]. We propose a new SSFSOD method to address this paradigm, which successfully embeds a teacher-student model into a meta learning-based SSOD framework. In particular, for each unlabeled image, we generate an easy-hard image pair by using strong data augmentation. Then we feed the easy image to the teacher model and the hard image into the student model. Previous studies [35, 40] in semi-supervised learning have shown that the teacher model's prediction of the easy image is significantly more accurate than the student model's prediction of the hard image. Therefore, we use the teacher model to generate reliable pseudo-labels to train the student model. The student model gradually updates the teacher model through Exponential Moving Average (EMA) [40], which can effectively alleviate the detrimental effects due to noisy pseudo-labels. Intuitively, the teacher model will teach it something confident like a humble teacher, while the student model will learn the hard image like an enthusiastic student. In addition, we propose a two-fold correlation-guided attention module, which consists of a correlation-guided spatial attention module and a channel attention module. The correlation-guided attention module can highlight the potential regions and emphasize useful information channels for the task-specific category. Through attention learning, the information of support images can be better used to guide RPN to generate task-specific region proposals. Experimental results demonstrate that the proposed method can improve the performance of few shot detectors.

In summary, our contributions are as follows: (1) To the best of our knowledge, this work is the first research of this challenging semi-supervised form of the underexplored problem of few-shot object detection. We formulate the SSFSOD problem and define benchmarks for evaluation by adapting from meta-learning based FSOD. (2) We develop a new method under the proposed paradigm. The proposed method embeds a teacher-student model into few-shot object detection based on a meta-learning framework for improving the robustness of the few-shot detector. (3) we employ a two-fold correlation-guided attention module, which can integrate task-specific information and enhance features by highlighting potential regions and informative channels, significantly improving the performance of few-shot detection. (4) We conduct extensive experiments on three popular benchmark datasets and demonstrate

that our method can successfully learn to leverage unlabeled examples and outperform purely supervised results. This work serves as a baseline for future research along this direction.

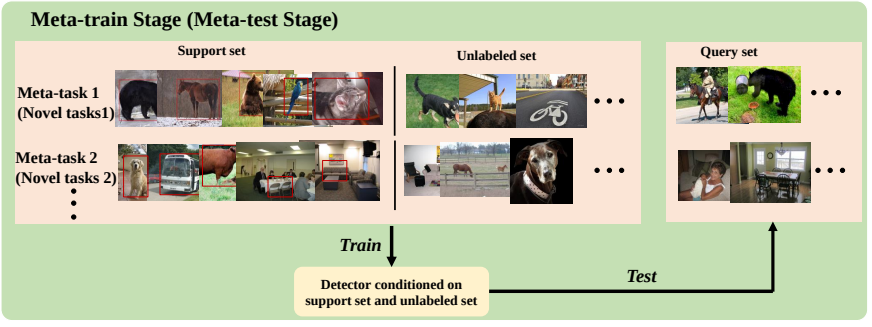

Figure 1: The overall framework of the proposed meta-learning based method for semi-supervised few shot object detection. **Meta-tasks** are sampled episodically from **source dataset** in meta-training stage, and **novel tasks** are constructed from **target dataset**. Moreover, the detector can use unlabeled images as auxiliary information in both the meta-test and meta-train stages.

2 Related work

2.1 Generic Object detection

Modern detectors based on deep learning could be further divided into two branches: two-stage detectors [13, 14] and single-detectors [2, 4, 26, 50]. Two-stage detectors first use an RPN to generate region proposals and then perform classification and fine-tuning the location of each region proposal, while single-stage detectors use densely placed anchor points as region proposals and directly predicts a label for them. Compared to a single-stage detector, Two-stage detectors have achieved state-of-the-art performance on many detection benchmarks [15, 23], but are generally slower. Single-stage detectors can achieve real-time inference speed, but the detection accuracy is usually not as good as two-stage detectors. The aforementioned methods rely on large-scale annotated datasets and are limited in many scenarios where the model has access to a few annotated training examples.

2.2 Semi-supervised object detection

In the standard semi-supervised target detection setting, the detector can use both labeled and unlabeled data during the training process. CSD [18] uses consistency constraints to force the predictions of the input image and its flipped version to be consistent for improving detection performance. Based on CSD, ISD [19] proposes a semi-supervised learning method for object detection based on interpolation, which directly applies interpolation regularization (IR) to object detection. STAC [37] combines self-training based on strong data augmentation and consistency regularization methods for semi-supervised object detection. Unbiased Teacher [27] uses the evolving student model to update the teacher model online

similarly to Mean Teacher [40]. Inspired by these methods [6, 27, 39], we embed the semi-supervised method into few-shot object detection based on a meta-learning framework to improve the performance of the few-shot detector.

2.3 Few-shot object detection

The previous work on few-shot detection can be divide into two paradigms: methods based on transfer learning and methods based on meta-learning. Methods based on transfer learning [6, 30, 41, 44] learn novel concepts through fine-tuning, while the methods based on meta-learning [11, 20, 45] adapt to new categories by extracting meta-level knowledge from learning various auxiliary tasks. Except for only a few recent works [11, 45], others cannot be directly applied to novel categories. Unlike the few examples of object detection [7] based on traditional semi-supervised learning, we perform semi-supervised object detection under meta-learning framework.

3 Method

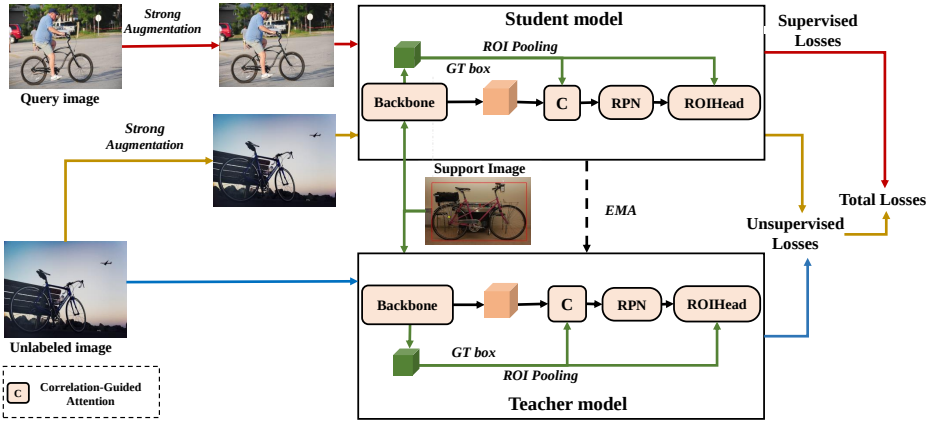

Figure 2: The framework of our proposed method. We re-formulate N-way-K-shot detection task into N binary detection tasks. The correlation-guide attention module would guide the RPN to generate specific-category proposals and filter out irrelevant proposals through the given support category. After the ROI pooling, we use the average feature across all the supports belonging to the same category as its support feature. The student model updates the teacher model via exponential mean average (EMA) manner. The final loss is the sum of supervised detection loss and unsupervised detection loss.

We build the proposed method upon the Faster-RCNN [32] which is a two-stage detector, first obtaining several regions of proposals through an RPN module and then extracting features from each region of the proposal to predict their categorical labels. The overview of our proposed method is shown in Figure 2. Before elaborating on the details of our method, we first describe our problem definition.

Problem formulation. Given four object detection datasets: D_{base}, D_{novel}, U_{base}, U_{novel}. D_{base} is large-scale dataset containing abundant annotated instances from each base class in

C_{base}. D_{novel} is a small-scale dataset with only K annotated instances from each novel class in C_{novel}. U_{base} and U_{novel} are unlabeled image datasets, which may contain objects of C_{base} and C_{novel} respectively. C_{novel} is disjoint from C_{base}. We perform semi-supervised few-shot object detection under a meta-learning framework containing two stages: meta-train and meta-test. During the meta-train stage, the detector needs to learn by a sequence of episodes, which contain a support set, query set and unlabeled image set. The support set and query set are sampled from D_{base}, and the unlabeled image set is sampled from U_{base}. During the meta-test stage, the detector requires to detect the target categories of objects for any query image in D_{novel}. If each support set contains N classes and K annotated instances for each class, the problem is called N-way-K-shot detection.

3.1 Correlation-guided attention

In few-shot object detection based on meta-learning, RPN filters out background and negative objects that do not belong to the support categories. However, without support instance information, RPN aimlessly generates many proposals of irrelevant objects, which burden subsequent classification tasks. The key is to embed the support information to the query feature and guide the RPN to produce specific-category proposals while suppressing other categories. In this work, we use two-fold correlation-guided attention to embed supporting information into the query feature.

Correlation-guided channel attention. Channel attention can make the feature more discriminative by recalibrating the channel feature response [1, 8, 16]. We first use channel correlation operations to obtain a similarity feature map between the support feature and the query feature by a channel-wise correlation and then learn channel attention from the similarity map. If a channel of the similarity map has a higher response, it helps distinguish task-specific objects and should be given a higher weight. Let $f_s \in R^{C\times H_s\times W_s}$ and $f_q \in R^{C\times H_q\times W_q}$ denote the support feature and query feature respectively. The feature map is given by:

$$f_c = f_s \star_c f_q \tag{1}$$

where $\star_c$ denotes channel-wise correlation and $f_c \in R^{C\times(H_q-H_s+1)\times(W_q-W_s+1)}$ is the correlation output. We use a global Maxpooling and Averagepooling on different channels to get a maxpool attention vector $f_c^{max} \in R^{C\times1\times1}$ and avgpool attention vector $f_c^{avg} \in R^{C\times1\times1}$. The two attention vectors are merged by using element-wise summation after passing a weight-shared one-layer perceptron (MLP). Then we use a sigmoid activation function to normalize their output range to [0, 1]. The process can be written by:

$$\alpha_c = \sigma\left(MLP\left(f_c^{max}\right) + MLP\left(f_c^{avg}\right)\right) \tag{2}$$

where $\alpha_c \in R^{C\times1\times1}$ and σ denote the channel-wise attention and the sigmoid function, respectively.

Correlation-guided spatial attention. In this work, we use a spatial attention module to focus on spatial regions that may contain target objects. We first reshape the support feature f_s to size $(H_s \times W_s) \times C \times 1 \times 1$. Then, we obtain the pixel-wise similarity by using the pixel-wise correlation between the support features and the query feature. The process can be expressed as:

$$f_p = f_s \star_p f_q \tag{3}$$

where $f_p \in R^{(H_sW_s)\times H_q\times W_q}$ and $\star_p$ denote the pixel-wise correlation operation and the correlation output, respectively. After pixel-wise correlation, an hourglass-like structure [29]

is used to learn a spatial attention map. Then, we normalize the output by using a sigmoid activation function to normalize the output:

$$\alpha_s = \sigma\left(H\left(f_p\right)\right) \tag{4}$$

where α_s and H denote the spatial attention map and the hourglass network.

After the two-fold attention module, we use a sequential multiplication to enhance the query feature with the channel-wise attention and the spatial attention maps. The process can be written as

$$f_q = f_q \otimes \alpha_c \otimes \alpha_s \tag{5}$$

where $\otimes$ is the broadcasting element-wise multiplication.

3.2 Supervised Loss

Our few-shot detector is a standard two-stage detector based on Faster R-CNN [32]. Its detection loss includes the RPN's loss and the ROI heads' loss. The supervised loss is written as:

$$L_s = L_{rpn} + L_{roi} \tag{6}$$

3.3 Unsupervised Loss

In this work, we use the soft label predicted by the detector as the training target of the unsupervised branch. Unsupervised loss is applied in both the detector's RPN (first stage) and ROI heads (second stage).

In the first stage, we apply the unsupervised loss to both RPN proposal classification and bounding box regression. Note that the data augmentation [36] we use does not cause the image geometry to change. Therefore, the teacher model and the student model share the same anchor set. The unsupervised loss for RPN's output is written as:

$$L_u^{rpn} = KL\left(t_{cls}^{rpn} \,||\, s_{cls}^{rpn}\right) + \left|\left|t_{reg}^{rpn} - s_{reg}^{rpn}\right|\right|_2 \tag{7}$$

where KL denotes the KL divergence. s_{cls}^{rpn} and s_{reg}^{rpn} denote the classification probability and bounding box regression output of student model RPN, and t_{cls}^{rpn} and t_{reg}^{rpn} are those of the teacher model RPN.

In the second stage, the RPN of the teacher model generates a set of candidate proposals, which is fed to the ROI heads of the teacher model and student model after a standard NMS [32] operation. The set of candidate proposals from the student model's RPN is not used for their ROI head training since the teacher model's proposals are often of higher quality than those from the student model. The unsupervised loss for the ROI heads is defined as:

$$L_u^{roi} = KL\left(t_{cls}^{roi} \,||\, s_{cls}^{roi}\right) + \left|\left|t_{reg}^{roi} - s_{reg}^{roi}\right|\right|_2 \tag{8}$$

where s_{cls}^{roi}, s_{reg}^{roi}, t_{cls}^{roi}, t_{reg}^{roi} denote the classification probabilities and all-class bounding box regression outputs by the student ROI head and teacher ROI head, respectively. The final unsupervised loss L_u is the sum of L_u^{roi} and L_u^{rpn}.

$$L_u = L_u^{roi} + L_u^{rpn} \tag{9}$$

3.4 Overall loss

During training, the final loss L is the sum of the supervised loss L_s and the unsupervised loss L_u,

$$L = L_s + \frac{n_u}{n_q}\lambda L_u \tag{10}$$

where n_u, n_q denote the numbers of unlabeled and query images,respectively. λ is a hyper-parameter which is set to 0.5 by default.

3.5 Teacher Model Updates via EMA

We use Exponential Moving Average (EMA) [40] to update the teacher model weights θ_t based on the student model weights θ_s. At each iteration, we have

$$\theta_t = \alpha\theta_t + (1-\alpha)\,\theta_s \tag{11}$$

where α denotes a hyperparameter. We set it to 0.999 by default. After each iteration, the teacher model only slightly updates itself from the student model. Teachers who are gradually updated are more resilient to sudden weight fluctuations of students. Even if the student model is fed the wrong label, the impact on the teacher model is mitigated by EMA.

4 Experiments

4.1 Dataset and Evaluation

We evaluate our approach on three detection datasets: MS COCO [22], Pascal VOC [10] and FSOD [11].

MS COCO. The MS COCO dataset [22] contains 80 classes, which are collected from various scenes on Flickr. Following the previous experimental settings [11, 45], we choose 20 classes (the same categories as the VOC dataset) as the novel (unseen) classes for evaluation and the remaining 60 classes as the base classes.

PASCAL VOC. The overall 20 categories in PASCAL VOC are divided into 15 base categories and 5 novel categories. Following the experiment settings as [45], we use VOC 07 and 12 train/val sets for training and VOC2007 test set for evaluation.

FSOD. The FSOD dataset contains 1000 categories with more than 60k images and 182k bounding boxes in total. For a fair comparison, we follow the setting in [11] to use 800 categories as base categories and 200 categories as novel categories.

Results on VOC. The results are reported in Table 2, where our proposed method surpasses all competitors under the supervised learning setting. When using unlabeled data, the performance of our model has been further improved.

For semi-supervised learning setting, we randomly sample images from the remaining dataset as an unlabeled set after sampling the support and query set in the meta-training and meta-test phases.

4.2 Experimental results

Results on MS COCO. Table 1 shows results reported in [11] and [45]. The all of existing methods perform few-shot detection in a supervised learning setting. As shown in Table 1, our proposed method outperforms the best SOTA model in the supervised learning setting.

Under the setting of 20-way-10-shot, the proposed method achieves the best results when using 40 unlabeled images for each mini-batch. While for the setting of 20-way-30-shot, the proposed method achieves the best results when using 20 unlabeled images. We observe that using unlabeled data can improve the model's performance in both settings.

Results on FSOD. Table 3 shows the results on FSOD. Under the supervised learning setting, the proposed method outperforms the existing methods. Our method achieved the best results with 20 unlabeled images per mini-batch. Compared with the existing method, the proposed method obtains an AP50 improvement of 2.4% and an AP75 improvement of 3.0%.

Table 1: The performance on MS COCO val set with 20-way novel classes. Best results are in bold.

Shot	Method	Backbone	AP	AP50	AP75
10-shot	Meta-Yolo [20]	ResNet-50	5.6	12.3	4.6
	FRCN+ft [17]	ResNet-50	1.3	4.2	0.4
	FRCN+ft-full [17]	ResNet-50	6.5	13.4	5.9
	MetaDet [33]	VGG-16	7.1	14.6	6.1
	Meta R-CNN [31]	ResNet-50	8.7	19.1	6.6
	Meta-RCNN [35]	ResNet-50	9.5	19.9	7.0
	A-RPN [11]	ResNet-50	11.1	20.4	10.4
	Ours (supervised model)	ResNet-50	12.0	22.1	11.8
	Ours (20 unlabeled images per batch)	ResNet-50	12.8	23.7	12.0
	Ours (40 unlabeled images per batch)	ResNet-50	**13.2**	**24.5**	**12.4**
30-shot	Meta-Yolo [20]	DarkNet-19	9.1	19.0	7.6
	FRCN+ft [17]	ResNet-50	1.5	4.8	0.5
	FRCN+ft-full [17]	ResNet-50	11.1	21.6	10.3
	MetaDet [33]	VGG-16	11.3	21.7	8.1
	Meta R-CNN [31]	ResNet-50	12.4	25.3	10.8
	Meta-RCNN [35]	ResNet-50	12.8	27.3	11.4
	A-RPN [11]	ResNet-50	13.2	27.4	12.2
	Ours (supervised model)	ResNet-50	14.1	28.4	13.9
	Ours (20 unlabeled images per batch)	ResNet-50	**15.2**	**29.6**	**14.5**
	Ours (30 unlabeled images per batch)	ResNet-50	15.1	29.1	14.3

Table 2: The mAP performance on Pascal VOC benchmark. All the models are evaluated with 5 ways on VOC2007 test set. Best results are in bold.

Method	Backbone	1-shot	2-shot	3-shot	5-shot	10-shot
YOLO-joint	ResNet-101	0.0	0.0	1.8	1.8	2.8
YOLO-ft	ResNet-101	3.2	6.5	6.4	7.5	12.3
YOLO-ft-full	ResNet-101	6.6	10.7	12.5	24.8	38.6
FRCN+joint	ResNet-101	2.7	3.1	4.3	11.8	29.0
FRCN+ft	ResNet-101	11.9	16.4	29.0	36.9	36.9
FRCN+ft-full	ResNet-101	13.8	19.6	32.8	41.5	45.6
LSTD [8]	ResNet-101	8.2	11.0	12.4	29.1	38.5
Meta-Yolo [20]	ResNet-101	14.8	15.5	26.7	33.9	47.2
MetaDet-YOLO [33]	VGG16	17.1	19.1	28.9	35.0	48.8
MetaDet-FRCN [33]	VGG16	18.9	20.6	30.2	36.8	49.6
Meta-CNN [31]	ResNet-101	19.9	25.5	35.0	45.7	51.5
Meta-RCNN [35]	ResNet-101	20.4	26.3	37.2	45.9	53.1
Ours (supervised learning)	ResNet-101	21.2	29.1	40.7	46.6	54.9
Ours (5 unlabeled images per batch)	ResNet-101	**22.1**	**29.9**	41.2	47.4	55.6
Ours (10 unlabeled images per batch)	ResNet-101	21.8	29.7	**41.5**	**47.6**	**55.9**

4.3 Ablation Study and Discussion

We conduct ablation experiments on the MS COCO dataset.

Effect of correlation-guided attention. The results of the three datasets show that our model with a two-fold correlated-guide attention module outperforms existing models in the

Table 3: Experimental results of our model on FSOD test set with 200-way 5-shot.

Method	AP50	AP75
FRCNN [32]	23.0	12.9
LSTD [6]	24.2	13.5
FSOD [11]	27.5	19.4
Ours (supervised learning setting)	28.5	21.2
Ours (10 unlabeled images per batch)	28.9	22.0
Ours (20 unlabeled images per batch)	**29.9**	**22.4**

supervised learning setting. In order to analyze the combined contribution of the channel and spatial modules, we show the results of using only channel attention (C_Att) and only using spatial attention (S_Att). The results in Table 4 show that S_Att achieves better performance than C_Att, and they both improve the performance. These results indicate that spatial attention and channel attention play an important role in our method, but spatial-wise attention is the more important one; this is reasonable because few-shot object detection requires rich spatial information from the support set. However, the combination of channel and spatial-wise attention leads to better performance.

Table 4: The performance (in %) with two-fold correlated-guide attention module on MS COCO. Best results are in bold.

Shot	Method	AP	AP50	AP75
10	S_Att	11.7	21.8	11.7
	C_Att	11.5	21.3	11.4
	S_Att + C_Att	**12.0**	**22.1**	**11.8**
30	S_Att	13.7	28.0	13.5
	C_Att	13.5	27.8	13.3
	S_Att + C_Att	**14.1**	**28.4**	**13.9**

Table 5: The performance (in %) with different number of unlabeled images on MS COCO. Best results are in bold.

Unlabeld	10-shot			30-shot		
	AP	AP50	AP75	AP	AP50	AP75
0	12.0	22.1	11.8	14.1	28.4	13.9
10	12.6	22.9	11.8	14.8	28.9	14.2
20	12.8	23.7	12.0	**15.2**	**29.6**	**14.5**
30	12.9	23.9	12.1	14.9	29.0	13.9
40	**13.2**	**24.5**	**12.4**	15.0	29.2	14.1
50	13.0	23.9	12.2	15.1	29.4	14.2

Number of unlabeled images. We explore the impact of different numbers of unlabeled images. Table 5 shows the results of using different numbers of unlabeled images on MS COCO. Our proposed method can achieve better performance with unlabeled images. The results begin to saturate after 40 unlabeled images for the 20-way-10-shot setting for 20 unlabeled images at the 20-way-30-shot setting. This shows that our method does not benefit from too much-unlabeled data. Overall, the results show that our method can effectively utilize unlabeled data.

Effect of the EMA. We examine the effects of using EMA training. We evaluate the model using various EMA rate α from 0.5 to 0.9999 and present the result in Table 6. Note that our model without EMA (w/o) is where the teacher and student model are shared-weights at each iteration. The student model and the teacher model are both updated by the supervised loss and the unsupervised loss, which is similar to the semi-supervised object detection method based on consistency constraints [18]. We observe that the model's performance improves with the increase of the EMA rate α. When the EMA rate α achieves 0.999, it performs the best performance. However, the performance begins to saturate after $\alpha = 0.999$. Table 6 also shows that our model with EMA is better than the model without EMA. The results suggest that EMA updates are crucial for our student-teacher model to work well. One possible explanation is that the model without EMA training suffers from an imbalanced pseudo-label

Table 6: The performance (in %) with the EMA training on MS COCO. Best results are in bold.

Shot	Model	AP	AP50	AP75
10	$\alpha = 0.5$	6.2	12.5	5.3
	$\alpha = 0.9$	10.7	20.4	7.2
	$\alpha = 0.99$	12.4	22.9	11.8
	$\alpha = 0.999$	**13.2**	**24.5**	**12.4**
	$\alpha = 0.9999$	13.0	23.7	12.1
	w/o EMA	12.7	22.9	11.8
30	$\alpha = 0.5$	9.4	19.7	8.1
	$\alpha = 0.9$	13.7	27.7	11.8
	$\alpha = 0.99$	14.3	28.1	13.7
	$\alpha = 0.999$	**15.2**	**29.6**	**14.5**
	$\alpha = 0.9999$	14.9	28.9	14.0
	w/o EMA	14.5	28.7	13.9

problem in the meta-learning framework. In few-shot object detection, most region proposals are from the background, and few are from task-specific classes. Therefore, when the model makes contradictory predictions for two corresponding region proposals (*e.g.*, background and the task-specific categorical label are 0 and 1, respectively), it has a higher chance of learning to predict 0 instead of 1. It is worth noting that this has not been identified as a serious problem under the supervised learning setting. This is mainly because ground-truth labels in training can easily suppress most background region proposals, so the model can focus on task-specific region proposals. But it is a different story in SSFSOD because we do not have highly reliable labels. Using the updated model of EMA alleviates the problem because the updated teacher of EMA can produce more accurate predictions than students by using easy images.

4.4 Training details

Our model was trained end-to-end using SGD on 4 GTX Titan X GPUs, with a batch size of 8 (for query images). We use a learning rate of 0.004 for the first 112,000 iterations and a learning rate of 0.0004 for the subsequent 8000 iterations. The short side of the query image and the unlabeled image is adjusted to 600 pixels; the longer side is capped at 1000. The support image is cropped around the target object with 16-pixel image context, zero-padded and then resized to a square image of 320 × 320. We follow the method in [36] to perform data augmentation for unlabeled images. Our implementation builds upon the Detectron2 framework [46].

5 Conclusion

In this work, we propose a novel semi-supervised few-shot object detection paradigm, where the detector can use an unlabeled image set as auxiliary information in each episode. Under this paradigm, we developed a semi-supervised few-shot object detection method, which use a teacher-student model to improve detection performance. In addition, we introduce a two-fold correlation-guided attention module to integrate task-specific information and enhance features by highlighting regions and informative channels, significantly improving the accuracy of few-shot object detection. Our proposed method achieves satisfactory performance on multiple object detection datasets.

Acknowledgement. This work was partially supported by the Academy of Finland under grant 331883 and the National Natural Science Foundation of China under Grant 61872379. The authors also wish to acknowledge CSC IT Center for Science, Finland, for computational resources.

References

[1] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, 2016.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[5] Cong Chen, Shouyang Dong, Ye Tian, Kunlin Cao, Li Liu, and Yuanhao Guo. Temporal self-ensembling teacher for semi-supervised object detection. *TMM*, 2021.

[6] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *AAAI*, 2018.

[7] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *TPAMI*, 2019.

[8] Fei Du, Peng Liu, Wei Zhao, and Xianglong Tang. Correlation-guided attention for corner detection based visual tracking. In *CVPR*, 2020.

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[11] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[13] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[17] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, 2019.

[18] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019.

[19] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. *arXiv preprint arXiv:2006.02158*, 2020.

[20] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019.

[21] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, 2019.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[24] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *IJCV*, 2019.

[25] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *IJCV*, 2020.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[27] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021.

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[30] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, 2020.

[31] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2016.

[33] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[34] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[35] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

[36] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, 2020.

[37] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[38] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020.

[39] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021.

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017.

[41] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020.

[42] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 2020.

[43] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *ICCV*, 2019.

[44] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *ECCV*, 2020.

[45] Xiongwei Wu, Doyen Sahoo, and Steven Hoi. Meta-rcnn: Meta learning for few-shot object detection. In *ACM MM*, 2020.

[46] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[47] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *ICCV*, 2019.

[48] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *CVPR*, 2020.

[49] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, 2020.

[50] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.