Few-Shot Object Detection via Baby Learning

Anh-Khoa Nguyen Vu[a,b,1], Nhat-Duy Nguyen[a,b,*,1], Khanh-Duy Nguyen[a,b], Vinh-Tiep Nguyen[a,b], Thanh Duc Ngo[a,b], Thanh-Toan Do[c] and Tam V. Nguyen[d]

[a] *University of Information Technology, Ho Chi Minh City, Vietnam*
[b] *Vietnam National University, Ho Chi Minh City, Vietnam*
[c] *Monash University, Clayton, VIC 3800, Australia*
[d] *University of Dayton, Dayton, OH 45469, United States*

ARTICLE INFO

Keywords:
Few-shot Object detection
Few-shot Learning
Baby Learning

ABSTRACT

Few-shot learning is proposed to overcome the problem of scarce training data in novel classes. Recently, few-shot learning has been well adopted in various computer vision tasks such as object recognition and object detection. However, the state-of-the-art (SOTA) methods have less attention to effectively reuse the information from previous stages. In this paper, we propose a new framework of few-shot learning for object detection. In particular, we adopt Baby Learning mechanism along with the multiple receptive fields to effectively utilize the former knowledge in novel domain. The proposed framework imitates the learning process of a baby through visual cues. The extensive experiments demonstrate the superiority of the proposed method over the SOTA methods on the benchmarks (improve average 7.0% on PASCAL VOC and 1.6% on MS COCO).

1. Introduction

Object detection has been a successful application domain of convolutional neural networks (CNNs). In literature, numerous works have been proposed such as Faster RCNN Ren, He, Girshick and Sun (2016), EfficientDet Tan, Pang and Le (2020), RetinaNet Lin, Goyal, Girshick, He and Dollár (2017b), YOLO Bochkovskiy, Wang and Liao (2020); Redmon and Farhadi (2018), and SSD Liu, Anguelov, Erhan, Szegedy, Reed, Fu and Berg (2016). These deep learning-based approaches primarily rely on abundant data in the training phase. However, this condition is not always true, especially when annotated data is extremely scarce. Given novel classes along with a few training samples, an object detection model trained on known classes needs to learn to detect novel class objects. This learning of these novel knowledge is commonly referred as few-shot object detection (FSOD).

Recent works in FSOD have been successful in overcoming the data scarcity by constructing meta-learning aids Kang, Liu, Wang, Yu, Feng and Darrell (2019); Wang, Ramanan and Hebert (2019); Yan, Chen, Xu, Wang, Liang and Lin (2019) or leveraging a fine-tuning technique Wang, Huang, Darrell, Gonzalez and Yu (2020). Of all those methods, the **t**wo-stage **f**ine-tuning **a**pproach (TFA) Wang et al. (2020) obtains the state-of-the-art (SOTA) performance in FSOD. Generally, the more visual shots a model learns, the more knowledge the model captures. However, Tab.1, which presents visual concepts between the different shots, gives a look at the performance of TFA (shown in Tab.1 for the original model and our reproduced model), the performance in a single shot learning is better than 2- or 3-shot learning. In addition, the knowledge quantification of 3-shot learning is worse than the one of 2-shot, as also shown in Tab.1.

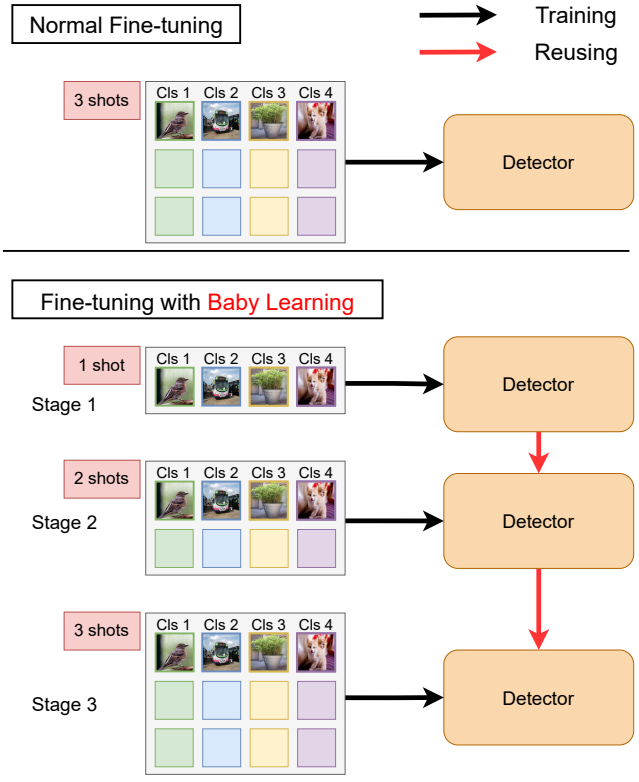

Figure 1: The overall difference of the learning scheme between normal fine-tuning and Baby Learning one. Our proposed model, FORD+BL (Baby Learning Multi-receptive Fields), adopts the baby learning mechanism along with the multiple receptive fields. Finally, FORD+BL significantly outperforms TFA.

Note that we follow Cheng *et al.* Cheng, Rao, Chen and Zhang (2020) to compute the knowledge quantification over the bounding box classifiers. The metric measures the

*Corresponding author
ORCID(s): 0000-0001-8566-273X (N. Nguyen)
[1] Equal Contribution

Method	Metric / Shot	1	2	3	5	10
TFA Wang et al. (2020)	mAP@50 ↑	39.8	36.1	44.7	55.7	56.0
TFA* Wang et al. (2020)	mAP@50 ↑	37.3	40.6	35.3	43.8	52.9
	Visual Concepts $\times 10^2$ ↑	19.3	22.3	21.7	23.6	25.3

Table 1
Motivation of Baby Learning in our work. Given very few shots such as 2 or 3 shots, their performance is always lower than 1 shot regardless of random or fixed instance in shots. *Our re-implementation with fixed shots.

discriminative power of features. Models with high visual and conceptual value should focus more on objects and gain better recognition abilities. Read Cheng et al. (2020) for more details. This abnormal problem shown in Tab.1 may occur due to a large variation on object appearances as a few samples that are used to train a few-shot model. On the contrary, humans, especially babies, have a perceptive ability of recognizing new concepts quickly by only being exposed to a very few samples. At first, their parent(s) shows a totally new object instance to form the initially visual recognition capability of the babies. Then, to further encourage strong visual development, the babies are rapidly taught with diverse instances. Since then the baby gradually improves his/her recognition capability to recognize the unseen instances. Note that this observation does not mean baby truly learns in this way from neuronscience perspective. Hence, we mimic this learning scheme to train the model with a totally new object at the very first time. This behaviour allows our model to reduce the variability of training data while the generalization of the detector is enhanced by being trained with more and more samples at the next times.

It is also worth noting that TFA adopts a two-stage fine-tuning technique which outperforms meta-based methods. In comparison to meta-learning works with a complex aid from the meta network, TFA simply fine-tunes only the last box predictor on novel classes while the rest is frozen during the fine-tuning stage. By freezing most of the layers in the network, TFA reuses the prior knowledge of base objects to predict novel ones. The simple approach takes a big gap compared to the previous works. However, they are not well-learned about the novel appearance that could be far different from the base domain due to freezing the most network. This makes the detector ignore the potential appearance of unseen classes or predict low-quality ones with low confident scores. To address this issue, we leverage the ability of multiple receptive fields to capture more spatial locations of an object in an image to its surrounding. This capability allow us to exploit the potential appearance in the base domain and use them in the novel domain effectively by fine-tuning the multiple receptive module. Therefore, we propose a new architecture which adopts the multi-receptive field named as **F**ew-sh**O**t with Multiple **R**eceptive Fiel**D** (FORD). In this paper, our contributions are three-fold:

- First, motivated from the early observations, we propose a novel and straightforward learning mechanism called ***B**aby **L**earning (BL)* in a way that imitates the learning process of a baby through visual cues to reinforce the development of visual recognition. Fig.1 illustrates the difference between of normal fine-tuning and **BL** approaches.
- Second, we propose a new architecture that leverages multi-receptive to capture the variant object appearance. Then, we fine-tune the new model to learn new shapes in the novel domain. Our model is named as **F**ew-sh**O**t with Multiple **R**eceptive Fiel**D** (FORD).
- Finally, we apply **BL** mechanism for **FORD** as a new framework for few-shot object detection learning. The extensive experiments demonstrate the superiority of the proposed method over state-of-the-art methods on benchmark datasets.

2. Related Work

Generic Object Detection. From the first introduction of RCNN Girshick, Donahue, Darrell and Malik (2014) in a series of RCNN family, object detection re-emerges with consecutively proposed methods Bochkovskiy et al. (2020); He, Gkioxari, Dollár and Girshick (2017); Karlinsky, Shtok, Harary, Schwartz, Aides, Feris, Giryes and Bronstein (2019); Liu et al. (2016); Redmon and Farhadi (2018); Ren et al. (2016); Tan et al. (2020) . These methods are then grouped into two genres such as one-stage and two-stage approaches. The one-stage approach aims to deal with object detection by proposal-free methods YOLO Bochkovskiy et al. (2020); Redmon and Farhadi (2018), RetinaNet Lin et al. (2017b), SSD Liu et al. (2016). While the two-stage approach includes methods such as RCNN family He et al. (2017); He, Zhang, Ren and Sun (2015); Ren et al. (2016), Efficientdet Tan et al. (2020) focusing on proposal-based algorithms Uijlings, Van De Sande, Gevers and Smeulders (2013). Most of the methods improve their performance by using informative characteristics on multi-scale feature maps. For example, YOLOv3 Redmon and Farhadi (2018) made its prediction based on three different scales of feature maps from Darknet53, while YOLOv4 Bochkovskiy et al. (2020) additionally chose the SPP block to enhance context features. SSD Liu et al. (2016) added multiple feature layers with different scales decreasing in size progressively after the backbone and deployed default boxes of different scales and aspect ratios. RetinaNet Lin et al. (2017b) used a newly proposed focal loss and attached FPN as its backbone to create a pyramid of feature maps. Similarly, EfficientDet Tan et al. (2020) was proposed with BiFPN for fusing multi-scale features.

Few-Shot Object Detection. FSOD refers to learning from just a few training examples per class. To date, there have been several works Kang et al. (2019); Yan et al. (2019); Wang et al. (2020); Fan, Zhuo, Tang and Tai (2020) focusing on FSOD. Two prior works Kang et al. (2019); Yan et al. (2019) mainly aim at tackling the problems of FSOD via meta-learning approaches that learn supportive information

from their meta learner to help models overcome the difficulties of the scarcity of data. Meta RCNN Yan et al. (2019) used labels of bounding boxes and segmented masks to train their meta network called Predictor-head Remodeling Network for inferring attention features. Feature Reweighting Kang et al. (2019) used a meta-model that takes mask areas of supportive objects formed by their associated object bounding box annotations to generate reweighting vectors for highlighting attention to each class. While Fan *et al.* Fan et al. (2020) exploited the advantages of support images from a massive FSOD dataset to generate significant results combined with their proposed network called Attention-RPN, Multi-Relation Detectors. The Attention-RPN directs the trained model where to look on the image for the task of object detection. Differently, Wang *et al.* Wang et al. (2020) simply adopted Faster RCNN and transferred massive knowledge from abundant data in the base model to fine-tune the novel one on few-shot data by freezing the whole network except for the fully connected layer for object classification. Through this simple straightforward mechanism, this model significantly improved few-shot performance without a complex pipeline of training the model.

However mentioned methods suffer from a drop in performance due to knowledge forgetting of base classes when trained on novel ones, Retentive R-CNN Fan, Ma, Li and Sun (2021), therefore, proposed Bias-Balanced RPN to debias the pretrained RPN and Re-detector to find few-shot class objects without forgetting previous knowledge.

Lately, there are newly proposed ideas for FSOD improvements by class correlation enhancement for discriminative power and multi-task learning with modified loss, namely Meta-DETR and DeFRCN. In 2021, based on Deformable DETR Zhu, Su, Lu, Li, Wang and Dai (2020), Meta-DETR Zhang, Luo, Cui and Lu (2021) utilized Correlational Aggregation Module (CAM) to aggregate query features with support classes. A highlight point of CAM is that CAM applies multiple support classes to query features simultaneously that allows the model to capture inter-class correlation during the training. In Qiao, Zhao, Li, Qiu, Wu and Zhang (2021), the authors presented DeFRCN employing Gradient Decoupled Layer to tackle FSOD from a multi-stage view. They also used Prototypical Calibration Block to decouple multiple tasks during the inference time.

Based on meta-learning appoarch, CME Bohao Li and Ye (2021) tried to create the classification feature space and optimize it via max-margin loss. The loss is calculated between intra-class and inter-class distance.

Spatial Pyramid Pooling. Exploiting the spatial details in an image has always been an essential component of the image analysis. However, very few works attached the spatial pyramid pooling for few-shot object detection. Most spatial attention models are often employed in segmentation or pose estimation. Several works used spatial pyramid pooling in their architecture like PSPNet Zhao, Shi, Qi, Wang and Jia (2017), Deeplab Chen, Zhu, Papandreou, Schroff and Adam (2018), UniPose Artacho and Savakis (2020) and DetectoRS Qiao, Chen and Yuille (2020). PSPNet Zhao et al. (2017) proposed a pyramid scene parsing network, which used 4 kernels with different sizes to extract feature maps and get global context. Deeplab Chen et al. (2018) used ASPP to improve performance and cost (storage, computation). In another work, instead of tackling segmentation, UniPose Artacho and Savakis (2020) proposed Water-fall Atrous Spatial Pooling (WASP) module and get significant performance in pose estimation. Likewise, DetectoRS Qiao et al. (2020) proposed Recursive Feature Pyramid to twice extract information from bottom-up and top-down in order to tackle both segmentation and detection tasks. Unlike the aforementioned methods, we mainly focus on object detection. Our model applies multi-receptive field to guide the model's attention to reliable potential bounding boxes.

3. Proposed Framework

In this section, we first summarize the traditional training phase in few-shot object detection. Then we refine this phase with BL. In terms of FSOD, we have two subsets of data to investigate in a detection dataset including base classes C_{base} and novel classes C_{novel}, where $C_{base} \cap C_{novel} = \infty$. The base classes are composed of abundant data with many instances of classes, denoted by $D_{base} = \{(I_i^{base}, y_i^{base})\}_{i=1}^{N_{base}}$, where $\{I_i^{base}\}$ represents input images from base class i, and $\{y_i^{base}\}$ denotes the associated object bounding box labels, and N_{base} is the number of base classes. Meanwhile, the novel classes solely contain a limited number of samples $D_{novel} = \{(I_i^{novel}, y_i^{novel})\}_{i=1}^{N_{novel}}$ that totally have K instances available per class. Here, $\{I_i^{novel}\}$, $\{y_i^{novel}\}$ are input images from novel class i, their associated object bounding box labels, and N_{novel} is the number of novel classes, respectively. Therefore, there are traditionally two stages to train a few-shot object detector. In the first stage, the detector is trained with the abundant data D_{base} from base classes C_{base} called base training. In the second stage or novel fine-tuning, the detector is continually deployed for training with the extension of novel classes C_{novel} and these novel classes only have a few labeled samples. For standard datasets such as PASCAL VOC and COCO, the novel set $S_{train} = D_{base} \cup D_{novel}$ for training is sampled in a balanced way to avoid problems of data imbalance and each class has the identical number of object annotations (K-shot, $K \in \{1, 2, 3, 5, 10\}$). We differentiate this setup from the few-shot detection scenario of a N-shot, M-way episode in ImageNet dataset for the FSOD task. This setup is derived from the few-shot classification literature.

3.1. Baby Learning

Shafto *et al.* Shafto, Conway, Field and Houston (2012) suggested that the visual learning in infancy is a sequential process. This non-linguistic learning ability plays a vital role to language development. We observe how the baby learns the new concept. At first parent(s) shows a totally new object instance to teach their baby about a new concept. At this point, an initial recognition capability about the

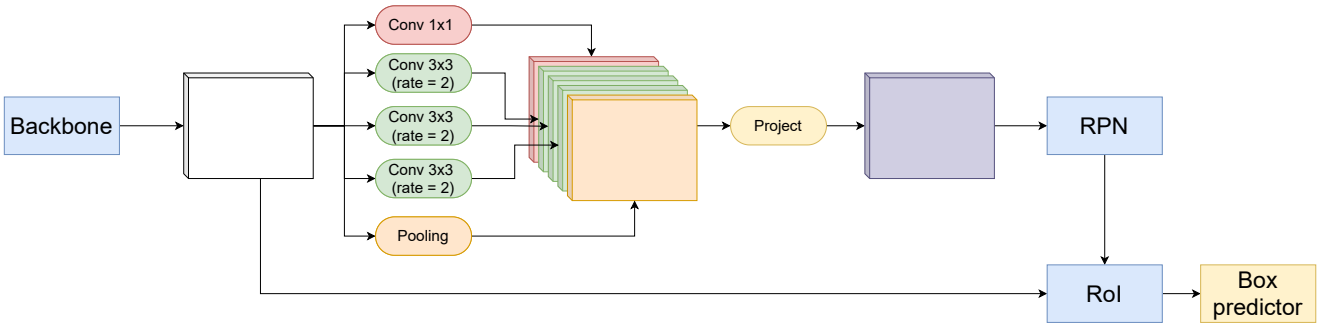

Figure 2: Overview architecture of FORD. In the base stage, the whole network is trained on base classes. In the novel stage, only the multiple receptive fields module and the box predictor are fine-tuned on a balanced data from both base and novel classes under baby learning mechanism. The blue rectangular boxes are the frozen modules in the novel stage.

concept emerges. Then, the baby is continuously taught with more visual instances to accumulate knowledge about the concept. Since then, the baby can gradually improve his/her recognition capability and recognize unseen instances. Motivated from the learning capability that a baby visually explores new stuff in the real world by learning from a few positive instances. These intuitive observations encourage us to propose a setup for training a few-shot detector called **Baby Learning**.

For an appearance of novel concepts, available works randomly initialize weights of the classification layer, then update learning weights via the novel training stage with few given samples per new class. This behavior unexpectedly gives a burden on the network when receiving new things in one time to learn with only a few available samples. Instead of doing a process that feeds all K instances one time per class ($K \geq 2$), we deploy a straightforward and simple BL that we first give a chance for our model to be familiar with a single shot in the very first time. Then the model is gradually trained with the greater number of shot instances. This practice is repeatedly done for each shot excluding one shot. Unlike normal fine-tuning , this "learning paradigm" inherits prior knowledge in the previously training times and thus the model comfortably exposes to more diverse and complicated samples afterwards. The conception can also be applied to many different tasks such as image segmentation, image generation and posture estimation. The following content is formulating BL approach:

We define $X = \{x_j\}_{j=1}^{n}$ as a set containing a number of instances in subshot of K-shot. X with n elements ($n \geq 2$) satisfies the following condition:

$$1 \leq x_j < x_{j+1} \leq K, \forall j \geq 2, \tag{1}$$

where x_j is the number of shots on D_{x_j} and $x_n = K$. The pretrained CNN on $D_{x_{j-1}}$ is trained on D_{x_j} continuously. If $j = 1$, we use the pretrained on D_{base}. The dataset D_{x_j} as follows.

$$D_{x_{j-1}} \subset D_{x_j} \subset D_{novel}, \forall j \geq 2 \tag{2}$$

The current benchmarks satisfy Eq.1 when we group the K-shot datasets (i.e., {1-shot, 3-shot} or {2-shot, 5-shot, 10-shot}). However, due to random shot generation, instances in 1-shot may be dissimilar in different K-shot (2, 3, 5). For that reason, the benchmarks cannot satisfy Eq.2 to implement BL. In this situation, to experiment our BL approach and compare with other methods on the same benchmarks, we split a K-shot dataset into subshots. For example, in 3-shot dataset benchmark, we get one and three instances to create D_{x_1} and D_{x_2}, respectively. The result of these steps is a new dataset having similar instances with 3-shot dataset and its X satisfies both Eq.1 and Eq.2 to implement BL approach, where $X = \{x_1, x_2\}$.

3.2. Baby Learning with Multiple Receptive Field

As studied in Gomez, Natu, Jeska, Barnett and Grill-Spector (2018), receptive fields (RFs) processing information in the visual field are a key property of human visual system neurons. Gomez *et al.* Gomez et al. (2018) found that multiple receptive fields were formed and developed early from childhood. Therefore, we aim to incorporate baby learning with multiple receptive fields. To this end, we first revisit TFA Wang et al. (2020). By freezing RPN in the fine-tuning stage, TFA Wang et al. (2020) leverages potential bounding boxes of base classes to predict the objects on novel classes. However, the diversity of the objects in the reality are very plentiful among classes and therefore the feature representations of novel objects could be far different from base objects. This makes the model miss potential bounding boxes or predict low-quality ones with low confident scores.

To deal with the problem, we enhance the localizability of the detection model in few-shot learning. We apply the multiple receptive field module to pay attention to more spatial information of an object in the image. As a result, FSOD algorithm can better recognize the base shapes and improve the generalizability in the new domain. Specifically, we apply multiple receptive fields by adopting the Atrous Spatial Pyramid Pooling (ASPP) Chen et al. (2018) in RPN branch of Faster RCNN Ren et al. (2016). The new model is named as **F**ew-sh**O**t with Multiple **R**eceptive Fiel**D** (FORD) that not only recognizes base objects better but also transfers the prior spatial knowledge to the novel domain effectively by fine-tuning multiple receptive fields module. The overview architecture of FORD is shown in Fig.2.

Method / Shot	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW† Kang et al. (2019)	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet† Wang et al. (2019)	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN† Yan et al. (2019)	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA Wang et al. (2020)	39.8	36.1	44.7	55.7	56.0	**23.5**	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
TFA* Wang et al. (2020)	37.3	40.6	35.3	43.8	52.9	23.4	27.2	35.4	33.2	39.5	28.5	37.2	42.6	48.4	48.9
Retentive R-CNN Fan et al. (2021)	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
FORD+BL	**46.3**	**54.2**	**49.9**	**56.3**	**61.8**	19.0	**30.8**	**38.4**	**39.3**	**47.3**	**36.4**	**46.5**	**45.4**	**53.2**	**55.8**
CME† Bohao Li and Ye (2021)	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
Meta-DETR† Zhang et al. (2021)	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6
Improvement from TFA	+9.0	+13.6	+14.6	+12.5	+8.9	-4.4	+3.6	+3.0	+6.1	+7.8	+7.9	+9.3	+2.8	+4.8	+6.9

Table 2
Few-shot detection performance (mAP) on the PASCAL VOC novel test set. The best performance is marked in boldfaced. *Our re-implementation with fixed shots. Methods in the grey color are not main comparison in this work. † denotes meta-learning approaches.

	AP@50			AP@75		
Method / Base set	1	2	3	1	2	3
TFA⋄ Wang et al. (2020)	81.1	80.7	81.6	**61.6**	61.8	62.2
FORD	**81.4**	**82.0**	**82.6**	60.8	**62.9**	**62.8**

Table 3
Detection performance (mAP@50 & mAP@75) on three different base sets of PASCAL VOC. ⋄ Our re-implementation.

3.3. Fine-tuning Implementation

In this subsection, we describe training stages that are applied to FORD in Fig.2. FORD uses the RFs module to gather more spatial information in base training. Then, we leverage the ability of the feature map with RFs to learn the positions from the novel object to its surrounding. In this way, we transfer the spatial information from the former domain to the current domain effectively.

Base model training. In the first stage, the model is trained on abundant data of base classes C_{base} to learn base knowledge which is suitable for the target domain. The joint loss function of Ren et al. (2016) is used during the optimization process.

Few-shot fine-tuning. In the first stage, our training data contains K shots per class for both base and novel classes. We assign randomly initialized weights for classes of the novel and reuse weights of base classes. During the time of training the model on the novel dataset, we jointly fine-tune both the box predictor and RFs module referred as jointly fine-tuning stage. Meanwhile, we freeze the rest of network weights to transfer the knowledge from the base to novel data.

4. Experiments and Discussion

In this section, we evaluate our method and compare it with previous works on the existing few-shot object detection benchmarks using PASCAL VOC 2007 Everingham, Van Gool, Williams, Winn and Zisserman (2010) and 2012 Everingham, Eslami, Van Gool, Williams, Winn and Zisserman (2015) and MS COCO Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick (2014). For the fair comparison, we follow the setup from Kang et al. (2019); Wang et al. (2020); Yan et al. (2019). In addition, for the ease of a comparison between the existing procedure on few-shot training and our proposed FORD+BL, we run all experiments by choosing fixed instances for each shot and re-produce the TFA results. In BL, with each shot except for 1-shot, FORD+BL is first trained with a fixed single instance, then with the whole instances of that shot.

4.1. Dataset and Settings

PASCAL VOC. PASCAL VOC is considered as a primary benchmark of object detection. The dataset comprises of 20 classes with two versions as VOC 2007 and VOC 2012. VOC 2007 trainval and VOC 2012 trainval sets are commonly used to train detectors and VOC 2007 test set is for testing in generic object detection. Regarding FSOD, both the trainval sets are separated into 3 certain splits in which 5 random classes are for the novel set and 15 remaining ones are keeping as the base set per split. Each novel class contains images with only K object instances that are available per class, where $K \in \{1, 2, 3, 5, 10\}$.

MS COCO. MS COCO is another challenging benchmark with a wide range of variation per class. In total, MS COCO has 80 classes. For FSOD experiments, COCO is split into two subsets, where the novel set consists of 20 classes overlapped with PASCAL VOC and 60 remaining ones are belong to the base set. 5000 images from the validation set, denoted as minival set, are used for evaluation while the left images in the training and validation sets are used for training. We follow previous works Kang et al. (2019); Wang et al. (2020); Yan et al. (2019) by setting $K = 10$ or $K = 30$ for MS COCO.

Implementation Details. In order to create individual shots from the given K shots, we crop object bounding boxes. In particular, we crop the target object with a random 20-pixel margin. Our model adopts Faster RCNN Ren et al. (2016) with ASPP Chen et al. (2018) and ResNet-101 backbone He, Zhang, Ren and Sun (2016) with Feature Pyramid Network Lin, Dollár, Girshick, He, Hariharan and Belongie (2017a). We use SGD optimizer with an initial learning rate of 0.004,

# shots	Method	Base classes															Novel classes						
		aero	bike	boat	bottle	car	cat	chair	table	dog	horse	person	plant	sheep	train	tv	mean	bird	bus	cow	mbike	sofa	mean	mAP
3	FSRW† Kang et al. (2019)	73.6	73.1	56.7	41.6	76.1	78.7	42.6	66.8	72.0	77.7	68.5	42.0	57.1	74.7	70.7	64.8	26.1	19.1	40.7	20.4	27.1	26.7	55.3
	Meta R-CNN† Yan et al. (2019)	67.6	70.5	59.8	50.0	75.7	81.4	44.9	57.7	76.3	74.9	76.9	34.7	58.7	74.7	67.8	64.8	30.1	44.6	50.8	38.8	10.7	35.0	57.3
	TFA* Wang et al. (2020)	86.3	87.5	72.8	70.8	87.8	86.4	62.3	77.3	82.7	83.5	86.0	51.7	78.0	86.6	82.0	78.7	18.7	31.3	34.4	50.5	41.8	35.3	68.0
	FORD+BL	87.9	79.8	70.3	73.8	88.2	88.2	55.7	73.2	85.3	86.3	86.7	48.9	81.0	87.0	79.1	78.1	37.3	64.2	55.9	51.8	40.5	49.9	71.1
10	FSRW† Kang et al. (2019)	65.3	73.5	54.7	39.5	75.7	81.1	35.3	62.5	72.8	78.8	68.6	41.5	59.2	76.2	69.2	63.6	30.0	62.7	43.2	60.6	39.6	47.2	59.5
	Meta R-CNN† Yan et al. (2019)	68.1	73.9	59.8	54.2	80.1	82.9	48.8	62.8	80.1	81.4	77.2	37.2	65.7	75.8	70.6	67.9	52.5	55.9	52.7	54.6	41.6	51.5	63.8
	TFA* Wang et al. (2020)	86.5	86.8	70.2	72.3	88.2	87.5	65.4	72.1	84.6	85.5	86.4	49.8	78.0	87.1	78.0	78.5	28.0	69.0	54.1	65.8	47.6	59.2	72.2
	FORD+BL	88.5	86.3	71.4	74.6	88.2	88.5	63.9	73.6	86.7	87.3	86.8	55.4	76.6	87.7	78.7	79.6	45.3	76.9	69.9	69.6	47.4	61.8	75.2

Table 4
AP and mAP on VOC test set for novel and base classes of novel set 1. Red and blue represent the best and the second best performance, respectively (Best viewed in color). *Our re-implementation with fixed shots. † denotes meta-learning approaches.

Shot	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
1	TFA* Wang et al. (2020)	**4.2**	**7.2**	**4.6**	**2.8**	**3.7**	**6.7**	**6.7**	**9.4**	**9.5**	**5.1**	**8.8**	14.0
	FORD	3.6	7.1	3.5	1.1	3.4	5.4	**6.7**	9.1	9.1	1.5	7.7	**15.3**
	Meta-DETR† Zhang et al. (2021)	7.5	12.5	7.7	-	-	-	-	-	-	-	-	-
3	TFA* Wang et al. (2020)	**7.1**	13.0	**7.0**	**3.6**	5.6	11.3	11.2	**17.3**	**17.5**	**7.9**	**15.8**	25.8
	FORD+BL	6.9	**14.3**	6.3	3.1	**6.6**	**11.6**	**11.3**	16.7	16.8	4.2	14.7	**28.5**
	Meta-DETR† Zhang et al. (2021)	13.5	21.7	14.0	-	-	-	-	-	-	-	-	-
5	Meta R-CNN† Yan et al. (2019)	3.5	9.9	1.2	1.2	3.9	5.8	-	-	-	-	-	-
	TFA* Wang et al. (2020)	**8.4**	16.1	**8.2**	**4.3**	7.3	13.0	12.8	20.1	20.3	**9.1**	**19.4**	28.8
	FORD+BL	8.2	**16.8**	7.3	3.6	**7.6**	**13.6**	**13.1**	**20.5**	**20.8**	6.4	18.5	**32.6**
	Meta-DETR† Zhang et al. (2021)	15.4	25.0	15.8	-	-	-	-	-	-	-	-	-
10	FSRW† Kang et al. (2019)	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
	MetaDet† Wang et al. (2019)	7.1	14.6	6.1	1	4.1	12.2	11.9	15.1	15.5	1.7	9.7	30.1
	Meta R-CNN† Yan et al. (2019)	8.7	19.1	6.6	2.3	7.7	14	12.6	17.8	17.9	7.8	15.6	27.2
	TFA Wang et al. (2020)	10.0	-	9.2	-	-	-	-	-	-	-	-	-
	TFA* Wang et al. (2020)	10.0	18.9	9.5	4.8	9.1	15.9	14.9	22.7	23.1	**10.2**	21.9	33.4
	FORD+BL	**11.2**	**22.5**	**10.2**	**5.2**	**10.1**	**18.2**	**15.9**	**24.8**	**25.3**	8.8	**23.8**	**38.3**
	CME† Bohao Li and Ye (2021)	15.1	24.6	16.4	4.6	16.6	26.0	16.3	22.6	22.8	6.6	24.7	39.7
	Meta-DETR† Zhang et al. (2021)	19.0	30.5	19.7	-	-	-	-	-	-	-	-	-
30	FSRW† Kang et al. (2019)	9.1	19	7.6	0.8	4.9	16.8	13.2	17.7	17.8	1.5	10.4	33.5
	MetaDet† Wang et al. (2019)	11.3	21.7	8.1	1.1	6.2	17.3	14.5	18.9	19.2	1.8	11.1	34.4
	Meta R-CNN† Yan et al. (2019)	12.4	25.3	10.8	2.8	11.6	19	15	21.4	21.7	8.6	20	32.1
	TFA Wang et al. (2020)	13.7	-	13.4	-	-	-	-	-	-	-	-	-
	TFA* Wang et al. (2020)	14.2	25.7	**14.3**	**5.9**	12.1	22.3	17.4	26.6	26.9	**10.0**	24.0	40.2
	FORD+BL	**14.8**	**28.9**	13.9	5.1	**14.5**	**23.3**	**18.5**	**28.8**	**29.3**	8.6	**27.2**	**44.6**
	CME† Bohao Li and Ye (2021)	16.9	28.0	17.8	4.6	18.0	29.2	17.5	23.8	24.0	6.0	24.6	42.5
	Meta-DETR† Zhang et al. (2021)	22.2	35.0	22.8	-	-	-	-	-	-	-	-	-

Table 5
Few-shot detection performance for the novel categories on COCO dataset. The '-' means the result is not reported in the original paper. * means our re-implementation with fixed shots. Methods in the grey color are not main comparison in this work. † denotes meta-learning approaches.

a mini-batch size of 4, momentum of 0.9 and the weight decay of 0.0001.

On PASCAL VOC, the base model is trained with 72,000 iterations and the learning rate is divided by 10 at 48,000 and 64,000 iterations. In the joint fine-tuning stage, we train our model for 3,500 and 500 iterations with learning rates of 0.004 and 0.0004, respectively. In the stages of BL, we trained model with the same configuration of the joint fine-tuning for 1, 2, 3, 5-shot and halved iterations of the first stages for 10-shot.

On MS COCO, the base model is trained for 180,000 iterations and the learning rate is divided by 10 at 120,000, 160,000 iterations. In the stages of fine-tuning with BL, the 10-shot dataset requires 8,000 iterations of the first stages and doubled for the last stage. In the 30-shot dataset, with $X_{30} = \{1, 5, 10, 20, 30\}$ defined in Eq.1, the model is trained for 8,000 iterations with 1-shot and 5-shot stages, then doubling the shots up to 30-shot with 32,000 iterations.

4.2. Results

We compare our approach with other SOTA methods Yan et al. (2019); Wang et al. (2019); Kang et al. (2019); Fan et al. (2021) and with the TFA baseline Wang et al. (2020) which is built upon Detectron2 Wu, Kirillov, Massa, Lo and Girshick (2019). For the ease of our evaluation, we first re-implement the experimental evaluation of TFA on the benchmarks (VOC and COCO) with fixed shots to compare with our model instead of random data in shots alike the origin in the paper, denoted TFA*. These experiments run

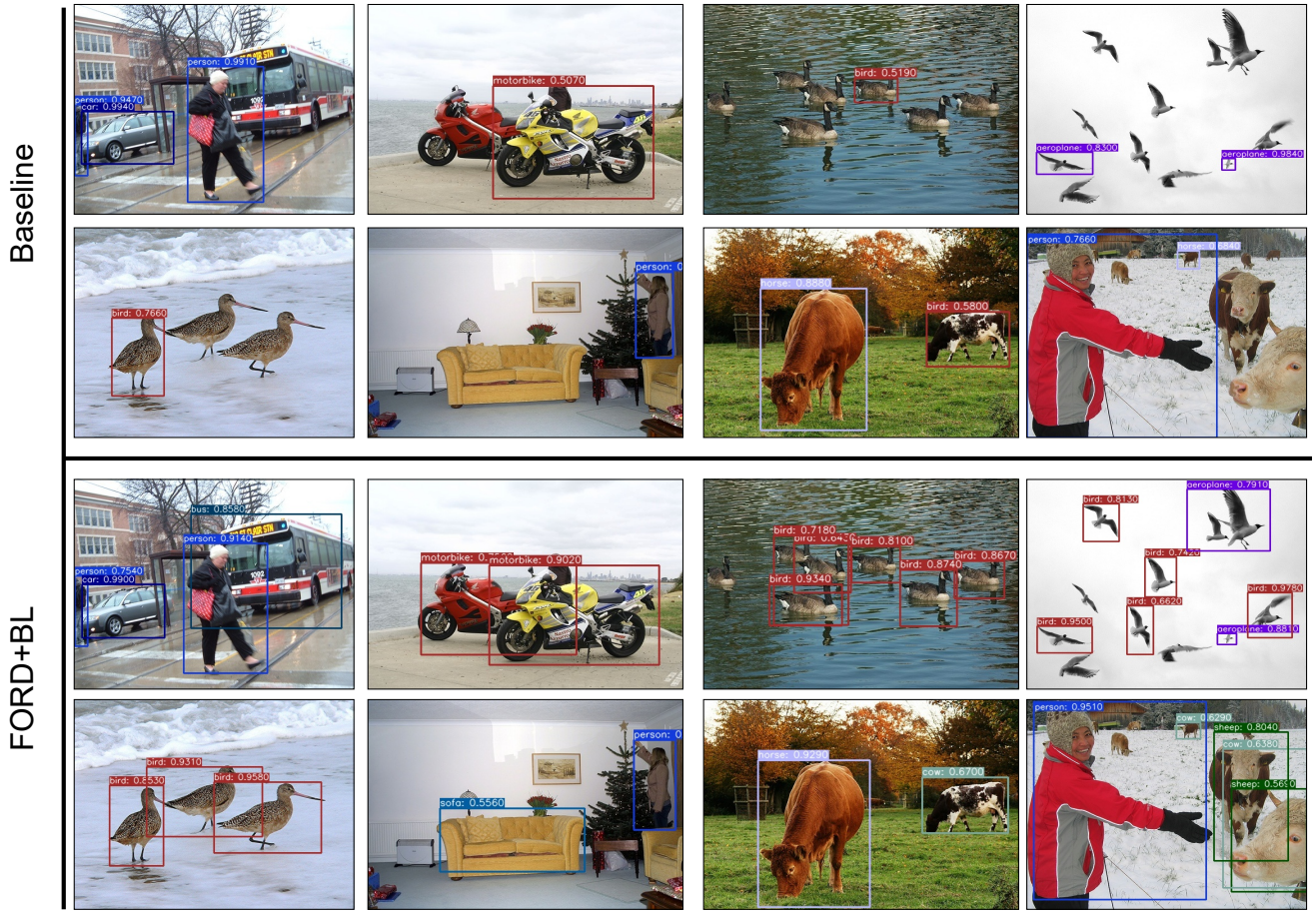

Figure 3: Qualitative 2-shot detection results on test set between TFA (top-2 rows) and FORD+BL (bottom-2 rows). Zoom in the figure for more visual details.

with the available source code and the original TFA use cosine similarity for the box classifier, which brings the best results for the TFA. While the model with RFs termed as FORD. We compare our approach FORD+BL formed by BL and RFs with FORD, the replicated TFA* and the origin TFA. In our experiments, FORD and FORD+BL do not use the cosine similarity to avoid the reliance on it.

PASCAL VOC. We first provide the experimental results of 3 novel sets on PASCAL VOC in Tab.2. Our proposed model FORD+BL receives SOTA on all 3 sets, even when labels are extremely scarce (1 or 2 shots). First, FORD+BL is greater than recently published methods such as Retentive R-CNN 3-5%; and significantly outperforms TFA about 5-7.5% except for the one shot, which BL is not applied for, in the novel set 2 lower than 4.4%. In case of the same setup between FORD+BL and TFA*, FORD+BL truly yields a remarkable performance on 3 splits. FORD+BL has a range of improvements in 3-11% in comparison with TFA* regardless of cosine similarity that have been used. Getting a closer look at the performance of TFA* on the novel set 1, 3-shot has a lower AP point than 1- and 2-shot. When FORD+BL is deployed, the 3-shot result is better than 1-shot. Specifically, the distances between TFA* and FORD+BL are approximately 13% AP for 2-shot and 3-shot. In the 1-shot, FORD demonstrates the efficiency of leveraging the RFs in FSOD. These results indicate that BL exceptionally reinforces FORD+BL with novel knowledge to well detect novel objects. In addition, FORD+BL along with RFs efficiently works on novel data by adapting spatial features on base data to exploit weak representation of unseen objects.

For more clear evidences of the capability for exploiting spatial features by multiple receptive fields, we also compare FORD with TFA on base categories only, shown in Tab.3. AP@50 means that predicted boxes and their corresponding ground truth have an overlap over 50%, similarly 75% for AP@75. Literally, an ideal model should run effectively when abundant data are given. Our proposal with RFs, FORD, outperforms in almost base sets except for AP@75 on the base set 1. This means FORD not only improves AP points on a normal but also stricter metric with the overlap ratio 0.75. This indicates that the model with RFs preferably captured better object representations.

We further compare the performance for each class in the novel set 1 on PASCAL VOC as shown in Tab.4. FORD+BL obtains the SOTA performance in both base

	2-shot		3-shot		5-shot		10-shot		
Subshot	1	2	1	3	1	5	1	5	10
Novel Set 1	41.8	54.2	43.7	49.9	44.2	56.3	42.6	54.6	61.8
Novel Set 2	15.3	30.8	15.7	38.4	16.4	39.3	12.5	39.1	47.3
Novel Set 3	28.8	46.5	27.3	45.4	28.3	53.2	26.4	46.8	55.8

Table 6
The performance of our proposed FORD+BL on PASCAL VOC splits. All shots are separated into subshots and then used to incrementally train our model. The last column in each shot indicates the final result whereas other columns are for intermediate results.

Methods	RFs	BL	1-shot	2-shot	3-shot	5-shot
TFA*			37.3	40.6	35.4	43.9
		✓	-	42.7	41.2	50.2
FORD	✓		**46.3**	50.4	43.7	48.4
	✓	✓	-	**54.2**	**49.9**	**56.3**

Table 7
Ablations of RFs and BL in Novel Set 1. *Re-implement with fixed shots and using settings in the origin paper.

and novel classes. Despite the slight increment in base classes, the improvement of FORD+BL in novel classes is significant. FORD+BL surpasses all baselines in terms of mAP. Compared with the second best in novel classes, FORD+BL leads a remarkable margin for most classes. Fig.3 shows qualitative 2-shot detection results between TFA and FORD+BL on base and novel classes. The two left columns show good cases indicating that our approach captures new objects in more shapes than the baseline. This is also shown in the failure cases of both TFA and FORD+BL in the two right columns of Fig.3. Though the two methods yield some false or miss-detection results, FORD+BL still productively leverages information from the previous stages and correctly detects more objects than TFA.

MS COCO. Tab.5 shows the comparative metrics on MS COCO dataset with $k = 10$ and $k = 30$. As shown in the table, FORD+BL significantly outperforms other methods including Retentive R-CNN and TFA with cosine similarity. In general, our framework achieves more 1.6% on average than TFA in both settings or than Retentive RCNN 1% on AP, *i.e.*, 10 and 30 shots. Our model gets significant performance on both AP and AR except for the small objects (32×32 pixels). In particular, our approach is about 4.6% higher for the large objects and 2.6% for the medium objects. In three common metrics (mAP, AP@50 and AP@75) our framework outperforms the other SOTA. It is worth noting that there is a large variation in object appearances in MS COCO. This is why the AP achieved in MS COCO is much lower than the one obtained in PASCAL VOC.

Extremely scarce data in MS COCO. We also conduct the experiments for 1, 3 and 5-shot in Tab. 5. The results of FORD are less improved or even worse when compared with TFA*.The reason is caused by the cosine layer which is applied to TFA* and is demonstrated to improve significant performance in the context of extremely scarce data. On the other hand, Meta-DETR gets incredible results in the same settings by using the meta-learning approach. We provide deep analyses in the Sec.4.3.

MS COCO to PASCAL. We evaluate our proposed framework on cross-dataset experiments. In this setup, we train all models on the MS COCO base classes and later fine-tune them on the 10-shot in PASCAL VOC. The results of TFA, FORD and FORD+BL are 38.7%, 43.9% and 47.5%, respectively. In general, they are worse than that when we use base classes in PASCAL dataset due to the large domain shift (i.e., number of classes or diversity of objects). The cross-dataset experiments of FSRW Kang et al. (2019) and Meta R-CNN Yan et al. (2019) are 32.3% and 37.4%, respectively. Our approach achieves 47.5% (compared to other methods we are about 10% higher in AP@50), which indicates that ours has high generality even in the context of cross-dataset.

BL Effectiveness. Tab.6 shows the performance of the integration of BL for our model on PASCAL VOC splits. As shown in the table, we note that fine-tuning with BL allows model to better learn new concepts when starting off from a single object to multiple objects. This opens a completely new approach in FSOD that a few-shot model could be progressively trained from simple to complex contexts to be familiar with diverse variants instead of directly feed them all with complicated instances once time. In this way, a few-shot object detection model could adapt to any new domains well, yet how many times of visual familiarity that a few-shot novice should be exposed prior to truly becoming a few-shot detector is still depending on the complexity of domains that BL is applied for.

Tab.7 and Tab.8 show further ablations on Novel set 1 of VOC to demonstrate the effectiveness of BL. Models with BL significantly achieve the superiority over ones without BL on all shots regardless of freezing or unfreezing modules. This means that with the aid of BL which learns from simple to complicated knowledge is very helpful for a few-shot learner so that they could tackle the diversity of variants in terms of a few available samples. The result also shows the consistency of changes in the variant diversity from very few to more available shots with and without BL paradigm. In addition, Tab.7 shows the performance difference between our model and the TFA baseline. Our models use RFs, which adapt spatial signal to predict novel appearances better than the TFA baseline regardless with/without BL. In the context of applying BL mechanism, the baseline and FORD have significant improvements for reusing the previous information of the novel domain (average improvements of 4.7 % for TFA* and 6.0% for FORD). Our model with BL all gains better performance than the TFA baseline with the average about 13.5%. Tab.8 provides an overall look on the contributions of different frozen modules that have less parameters but helps to describe crucial

Frozen module			BL	Novel Set 1			
RPN	ASPP	Project		1-shot	2-shot	3-shot	5-shot
✓				46.3	50.4	43.7	48.4
			✓	-	**54.2**	**49.9**	**56.3**
	✓			**46.4**	49.8	42.2	49.0
			✓	-	53.5	48.6	54.0
✓	✓			46.3	51.6	43.7	48.9
			✓	-	54.3	49.5	54.2
	✓	✓		46.0	49.2	40.8	48.8
			✓	-	53.9	48.4	54.7
✓	✓	✓		45.3	47.7	41.3	49.7
			✓	-	54.1	48.8	55.0

Table 8
Ablations of freezing modules with/without BL.

features during learning phase. The models with unfreezing the project layer gets the higher performance than others. Adding more frozen modules reduce the AP about 1-4%. This implies that when modules are frozen, they suffer from adapt to novel knowledge due to the domination of base data on the previous stage. Note that ASPP and the project layer are used to create the feature maps with multiple receptive fields.

4.3. Compare with meta-learning approaches

Previous works such as Kang et al. (2019);Yan et al. (2019); Wang et al. (2019) have demonstrated the benefits of the meta-learning approach when compared to the fully supervised-based models such as Ren et al. (2016); Redmon and Farhadi (2018) in the term of FSOD. However, latter works in 2020, our baseline architecture TFA Wang et al. (2020), for example, is fully based on Faster RCNN Ren et al. (2016) coming with a proposal which freezes the modules in the *few-shot fine-tuning* phase to get great superiority about 2-20% over former meta-learning approaches. Therefore, FORD+BL improves the TFA and enhances the gap to 3-34% when compared to above meta-learning methods.

In 2021, meta-learning based-methods bounce back and achieve outstanding results in FSOD, especially CME Bohao Li and Ye (2021) and Meta-DETR Zhang et al. (2021). On Pascal VOC, CME detector achieves comparative results with FORD+BL. While Meta-DETR demonstrates a great improvement over FORD+BL 4-10%. On COCO, CME achieves 2-5% compared to our work and Meta-DETR similarly outperforms CME by 4-6%. Both methods are designed to create effective feature space for learning discriminative features of novel classes from base model training. Concretely, CME Bohao Li and Ye (2021) designs a max-margin loss with an aim to optimize feature space partition. On the other hand, meta-DETR Zhang et al. (2021) exploits the inter-class correlation to enhance the generalization of the model.

It is worth noting that our method is based on Faster RCNN architecture Ren et al. (2016) with the fine-tuning approach similar to our baseline and **does not** use meta-learning techniques to produce the base model or the final prediction for novel classes. Besides, our proposed method can combine with meta-learning approaches to get potential results. Finally, with the way of BL mechanism, we can apply this learning to FSOD methods in order to further improve the performance by reducing the difficulty of the novel appearance and tackling the data scarcity.

4.4. Open Issues

One of the main challenges is to effectively leverage BL mechanism so that gains better performance and addresses the variant objects by differentiating a bad or good sample at the very first times. In addition, we have to clarify how many times of visual familiarity that a few-shot novice should be exposed prior to truly becoming a few-shot detector.

Another critical issue that has not been addressed yet is how to enhance the generalization of the model when novel and base objects all occur. We need to reduce the domination of base classes on novel domain with few samples while still remaining performance. Finally, it is vital to clarify the advantages of BL ability to apply it for other tasks.

5. Conclusion

We have presented a new framework dubbed FORD+BL for few-shot object detection that adopts the baby learning mechanism along with the multiple receptive fields. We first proposed the straightforward BL that benefits the model training. BL learns from diverse instances of novel classes by first being familiar with a single instance and then more visual variants. Meanwhile, multiple receptive fields allow the model to work well and overcome the data scarcity by only fine-tuning the project layer. Finally, our proposed model achieved the superiority over state-of-the-art methods on benchmark datasets.

In the future, we improve the proposed FORD+BL framework. We plan to explore the impact of object's size or occlusion in FSOD and FORD+BL. We notice the performance of the model on novel classes is affected by training setup on base classes. Hence, we aim to quantify the quality of the feature maps on base classes that is good for novel classes. FORD+BL demonstrates its superior performance on few-shot learning, however, we just fix instances for the fine-tuning stage and ignore the well-presented instances as representatives for a specific class, which is worth investigating in the future.

Acknowledgement

This research is funded by National Science Foundation (NSF) under Grant No. 2025234.

References

Artacho, B., Savakis, A., 2020. Unipose: Unified human pose estimation in single images and videos, in: CVPR.

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 .

Bohao Li, Boyu Yang, C.L.F.L.R.J., Ye, Q., 2021. Beyond max-margin: Class margin equilibrium for few-shot object detection, in: CVPR.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV.

Cheng, X., Rao, Z., Chen, Y., Zhang, Q., 2020. Explaining knowledge distillation by quantifying the knowledge, in: CVPR.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88.

Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W., 2020. Few-shot object detection with attention-rpn and multi-relation detector, in: CVPR.

Fan, Z., Ma, Y., Li, Z., Sun, J., 2021. Generalized few-shot object detection without forgetting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4527–4536.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR.

Gomez, J., Natu, V., Jeska, B., Barnett, M., Grill-Spector, K., 2018. Development differentially sculpts receptive fields across early and high-level human visual cortex. Nature communications 9.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: ICCV.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition, in: TPAMI.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T., 2019. Few-shot object detection via feature reweighting, in: ICCV.

Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M., 2019. Repmet: Representative-based metric learning for classification and few-shot object detection, in: CVPR.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection, in: CVPR.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection, in: ICCV.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: ECCV.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: ECCV, Springer.

Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., 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), pp. 8681–8690.

Qiao, S., Chen, L.C., Yuille, A., 2020. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. arXiv preprint arXiv:2006.02334 .

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks, in: TPAMI.

Shafto, C.L., Conway, C.M., Field, S.L., Houston, D.M., 2012. Visual sequence learning in infancy: Domain-general and domain-specific associations with language. Infancy Journal 17.

Tan, M., Pang, R., Le, Q.V., 2020. Efficientdet: Scalable and efficient object detection, in: CVPR.

Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. International Journal of Computer Vision 104.

Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F., 2020. Frustratingly simple few-shot object detection, in: ICML.

Wang, Y.X., Ramanan, D., Hebert, M., 2019. Meta-learning to detect rare objects, in: ICCV.

Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L., 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning, in: ICCV.

Zhang, G., Luo, Z., Cui, K., Lu, S., 2021. Meta-detr: Image-level few-shot object detection with inter-class correlation exploitation. arXiv preprint arXiv:2103.11731 .

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: CVPR.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 .