DF-FSOD: A Novel Approach for Few-shot Object Detection via Distinguished Features

Anh-Khoa Nguyen Vu[1,2], Thanh-Danh Nguyen[1,2], Vinh-Tiep Nguyen[1,2], and Thanh Duc Ngo[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{khoanva, danhnt, tiepnv, thanhnd}@uit.edu.vn

***Abstract*—Few-shot object detection (FSOD) is a challenging task in which detectors are trained to recognize unseen objects with limited training data. The majority of existing methods are evaluated on the benchmarks built with a fixed quantity of base and novel classes categories. To be specific, the number of base classes is larger than the novel ones. This positively affects the performance evaluated on novel data. However, there are not many works focusing on the effect of such dominated categories on the performance of FSOD models. In this paper, we investigate the efficiency of the detectors in different ratios of base and novel categories in the novel phase. Based on our findings of the affection between base and novel classes, we present a new approach: Distinguished Features for FSOD (DF-FSOD), which encourages the detector to learn distinguished features to capture novel objects via base-class expansion better. In the end, our proposed method outperforms average 4% AP@50 on PASCAL VOC compared to the previous works on the unseen classes when extremely scare labeled data.**

***Index Terms*—few-shot learning, object detection, distinguished features**

I. Introduction

In the context of computer vision, deep learning based approach achieves significant performance in various tasks such as classification [1]–[3], human pose estimation [4]–[6], and detection [7]–[9]. The state-of-the-art (SOTA) methods to solve the above problems mostly belongs to the approach of fully supervised learning in which models are trained with abundant data to get well-performance. However, the annotated datasets [10]–[12] are rare and the process of creating them is high-cost. In reality, human vision accurately recognizes the new objects with a few provided samples. The nerve cells quickly learn the new concepts thanks to leveraging the prior knowledge from the surroundings. From this observation, we teach the machine to imitate the ability of human vision to capture new things (so-called few-shot learning).

The objective of few-shot learning is to resolve a problem within limited annotated data. The pioneer works propose approaches in classification. Recently, few-shot learning adapts to segmentation [13], [14] and detection [15]–[18]. In generic object detection, many feature learning-based approaches such as Fast RCNN [7], Faster RCNN [8], EfficientDet [9], RetinaNet [19], YOLO [20]–[22], SSD [23] require a large amount of training data to recognize and localize the objects. Differently, in the few-shot object detection (FSOD), the traditional approaches leverage the features generated in seen (base) classes with plentiful training data to detect the unseen (novel) objects with just a few instances.

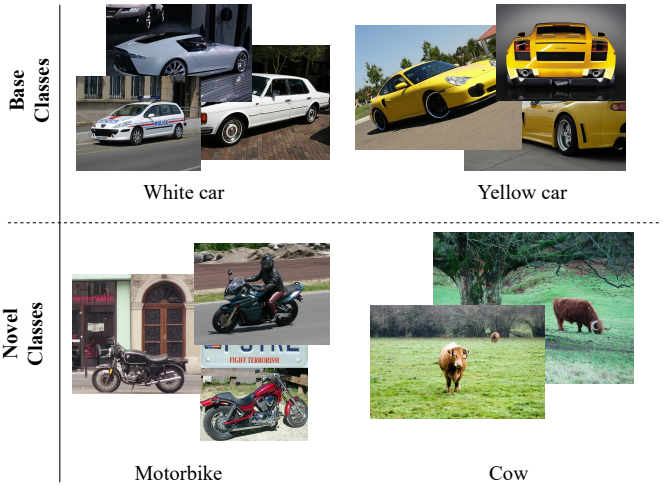

Fig. 1. Toy example of our approach for few-shot detection learning. For 3-shot dataset, we increase twice the number of base classes ("Car" class) while the number of novel classes ("Motorbike" and "Cow" class) is kept fixed as traditional settings. The "Car" class can be divided into two subclasses, "**White** car" and "**Yellow** car", to encourage the CNNs to learn distinguished features.

Most of FSOD methods [15]–[18], [24] are based on the two-stage fine-tuning approach to transfer knowledge from the base data to novel data. To be specific, Meta-RCNN [15] and Feature Reweighting [16] use an external branch called meta-learner to weight the features. The meta-learner creates the attention vectors for each category by taking the annotations as inputs, which is combined with feature maps from the backbone. In 2020, Wang *et al.* introduced a **t**wo-stage **f**ine-**t**uning approach (TFA [24]) using information from abundant data to learn the features of few shots by freezing almost the entire network. On the other hand, Fan *et al.* [18] uses weighted feature maps to calculate the relation between the query and support image. They also provide practical experiments to indicate that the performance of FSOD methods enhances when increasing the class diversity. However, they do not explain the results, and there are not many works investigating the effective use of the number of classes in the source domain (base classes).

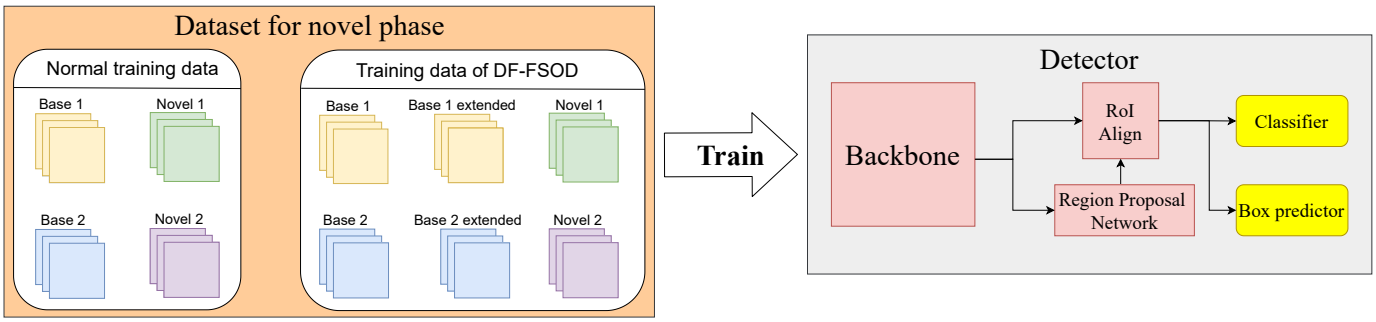

Fig. 2. Overview of our approach. The left rectangle describes the two ways to set up a dataset for the novel phase. The right rectangle describes the training of the detector in the novel phase. The red rectangular boxes are frozen, and the yellow rectangular boxes are fine-tuned in novel class.

In this work, we investigate the performance of detectors in FSOD when changing the number of classes. We apply the potential approaches to exploit the prior knowledge and encourage the models to learn distinguished features on novel classes during the fixed base training. Relied on our findings of the affection between base and novel classes in the novel fine-tuning phase, we observe that increasing the diversity of base classes in the novel phase (see Fig. 1) helps the model learn expensive features. We present a new approach for FSOD which allows the detector to localize and classify novel objects better. Our approach (shown in Fig. 2) promotes the pretrained detectors to learn the kernels that create discriminative visual information among similarity classes. Besides, the kernels also learn the novel concepts of unseen classes that force the models to effectively utilize kernels to generate useful features for both novel classes and extended base classes. Finally, our proposed approach creates detectors outperforming the previous ones on the unseen classes.

In this article, we propose three main contributions:

- Firstly, we investigate the performance of FSOD detectors on novel classes. There are works to deal with FSOD but less attractive to effectively exploit the power of the number of classes to improve the performance of algorithms on limited data.
- Secondly, we present a novel approach for FSOD which improves the performance of detectors in unseen objects. The key idea is expanding the base classes to encourage the model to leverage prior knowledge to design the distinguished features and improve performance in unseen classes. Our approach is named after **D**istinguished **F**eatures for FSOD (DF-FSOD).
- Finally, we provide practical experiments to demonstrate our methods. Specifically, the models based on our approach outperform others (average 4.0% AP@50 on PASCAL VOC 1-shot datasets). We also present the visualization to compare our results with the TFA [24] in Fig. 3.

II. Related Work

Object Detection. Object detection is a fundamental problem in computer vision. Deep CNN-based methods may be divided into two main approaches: one-stage and two-stage detectors. The one-stage approach attempts to predict bounding boxes directly, and scores of objects without an intermediate step such as YOLO [20]–[22], SSD [23], RetinaNet [19]. In contrast, the two-stage approach uses a buffer step (region proposal network [8], selective search [25]) to generate region of interests (ROI) and uses them to classify and regresses the location of objects. The classical methods of this branch are RCNN series [7], [8], [26], [27]. RCNN [26] feeds approximately 2000 region proposals generated by Selective Search [25] to a pretrained-CNN model to classify. Different RCNN, Fast RCNN [7] applies RoI pooling on feature maps to extract region proposals effectively. Faster RCNN [8] improves the performance of previous versions by using the region proposal network.

Few-Shot Object Detection. FSOD aims to learn a few instances to localize and recognize the new objects. There are many works [15], [16], [18], [24] to tackle this issue. Meta R-CNN [15], and Feature Reweighting [16] use the feature re-weighting mechanism to focus on each category with the support of meta-learner to utilize the support images or annotations as inputs. Meta R-CNN [15] is based on the architecture of RCNN [26] and leverages mask/box annotations to infer attentive vectors for each class. On the other hand, Feature Reweighting [16] is built on YOLOv2 [20] architecture and takes the mask of objects to create the reweighing vectors for each class. Differently, Fan *et al.* [18] applies the features effectively from Attention RPN module to support Multi-Relation Detector measuring the similarity between the query and the support objects. On the contrary, TFA [24] simply utilizes Faster RCNN [8] to leverage solid knowledge from abundant base data to fine-tune the novel few-shot data with freezing most layers in the network. However, the above methods simply use base classes to create base knowledge and maintain the performance of the detector in novel phase. They do not research the way to exploit base data when learning the new concepts effectively.

III. Effectively Distinguished Features

The process of constructing standard datasets is tough and costly. To deal with the problem, few-shot learning is introduced as a potential highly effective solution. However, there have not been many works for FSOD [15]–[18], [24] and the

TABLE I
FEW-SHOT DETECTION PERFORMANCE (AP@50) ON THE PASCAL VOC NOVEL TEST SET. *OUR RE-IMPLEMENTATION.

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
TFA [24] w/fc*	34.5	38.5	35.4	43.7	**52.5**	21.0	**27.2**	**33.0**	**33.4**	38.4	25.0	36.0	40.1	48.1	48.7
+ DF-FSOD	**37.9**	**39.5**	**36.0**	**45.0**	52.3	**21.3**	26.1	31.8	33.2	**39.1**	**30.0**	**38.0**	**41.4**	**47.4**	**48.8**
TFA [24] w/cos*	35.9	37.7	35.1	43.8	51.8	22.6	**26.2**	**34.3**	32.5	38.6	21.3	38.2	40.6	48.0	48.5
+ DF-FSOD	**40.6**	**39.7**	**35.3**	**45.9**	**52.4**	**24.3**	24.6	33.1	**34.1**	**38.9**	**30.4**	**40.2**	**42.1**	**48.8**	**49.0**

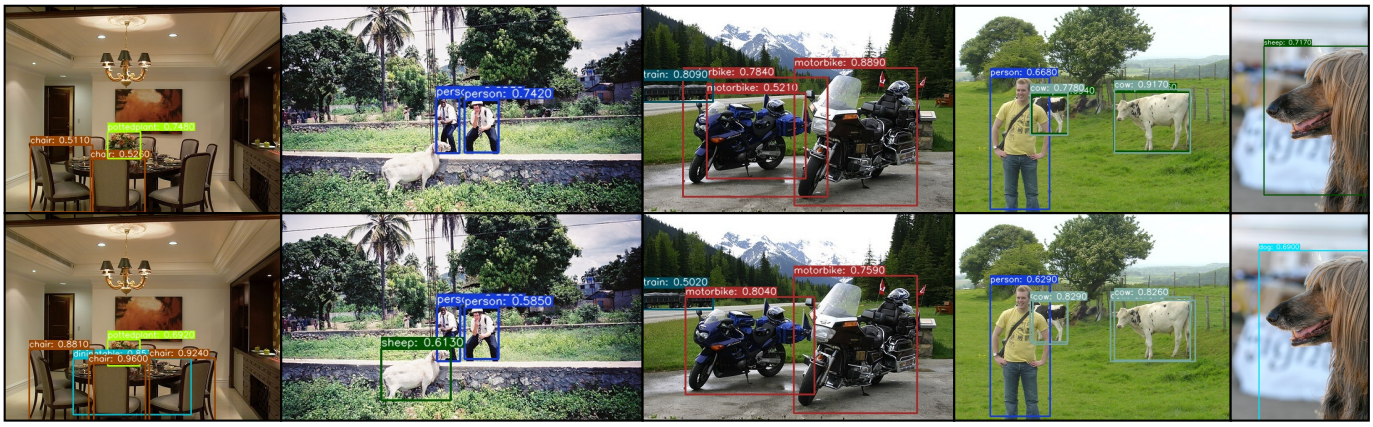

Fig. 3. Illustration of our results on 1-shot dataset in the novel third set of PASCAL VOC. The first row is the results of baseline. The second row is the results of our approach.

community is notably less interested in efficient data exploiting to improve model performance. Fan *et al.* [18] performed several experiments and indicated that increasing diversity in the data is necessary for FSOD without any explanation. In this paper, we have conducted research to effectively exploit the existing data in the source domain (base classes C_{base}) to improve the performance of detectors in the target domain (novel classes C_{novel}).

The weights are learned during the training of the algorithm to produce discriminative information. However, they also make ambiguous feature maps. Therefore, the model that generates a lot of discriminative information makes accurate predictions. In FSOD, the number of instances of novel classes is insufficient to generate useful features for object classification, which may be why FSOD methods [15]–[18], [24] are often good at localizing but bad at classifying novel objects. We observe that expanding the number of classes in the base classes C_{basel} encourages the model to distinguish features between objects of similar classes effectively (i.e., husky dog and pug dog). These effectively distinguished features support methods making the correct predictions of novel classes C_{novel} with high similarity (like a cat). We dub the approach appropriating distinguished features for few-shot object detection as **DF-FSOD**. To clarify, we formulate the problem in the following section.

IV. FEW-SHOT OBJECT DETECTION FORMULATION

This formulation is based on the few-shot object detection setting created by Feature Reweighting [16] on PASCAL VOC. There are two subsets of data to investigate the performance of FSOD consisting of base classes C_{base} and novel classes C_{novel}, where $C_{base} \cap C_{novel} = \emptyset$. The number of base classes C_{base} and novel classes C_{novel} are M_{base}, M_{novel}, respectively. The base dataset $D_{base} = \{(x_i^{base}, y_i^{base})\}_{i=1}^{N_{base}}$ have many instances per base class and the novel dataset $D_{novel} = \{(x_i^{novel}, y_i^{novel})\}_{i=1}^{N_{novel}}$ have K instances per novel class, where N_{base} and N_{novel} are a number of images in D_{base} and D_{novel}, respectively. In the term of few-shot learning, the CNN model uses massive knowledge from base dataset to understand the concept features in the novel dataset ($N_{base} \gg N_{novel}$ and $M_{base} \gg M_{novel}$). Therefore, there are pioneer methods that ordinarily apply two-stage approach to train the detectors in D_{novel} to predict both base and novel objects. In the first phase, the detectors are trained in D_{base} to learn extensive features of the objects in C_{base}, which is called base training. In the second phase or novel fine-tuning, the pretrained detectors are learned in D_{novel} to recycle solid knowledge from base training to capture novel objects in C_{novel} with K shots of objects ($K \in \{1, 2, 3, 5, 10\}$).

A. Constraint investigation

In this section, we formulate the constraints to investigate the new approach having comparable results. To compare

TABLE II
AP@50 AND MEAN ON PASCAL VOC TEST SET FOR NOVEL AND BASE CLASSES OF THIRD SET. *OUR RE-IMPLEMENTATION.

Shot	Methods	Novel Classes						Base Classes														
		Boat	Cat	Mbike	Sheep	Sofa	Mean	Aero	Bike	Bird	Bottle	Bus	Car	Chair	Cow	Table	Dog	Horse	Person	Plant	Train	Tv	Mean
1	TFA [24] w/fc*	**14.6**	**27.5**	45.6	25.0	11.7	25.0	**84.7**	**86.0**	**76.8**	**73.5**	**83.7**	**87.5**	**58.6**	**77.2**	**49.5**	**80.5**	**87.8**	**77.6**	**52.0**	**86.2**	**74.6**	**75.7**
	+ DF-FSOD	14.5	26.6	**56.6**	**40.4**	**11.9**	**30.0**	71.7	74.7	71.1	61.1	74.1	75.3	51.2	69	67.2	69.6	76.4	68.4	48.5	78.6	75.1	68.8
	TFA [24] w/cos*	13.0	19.3	43.1	23.4	7.6	21.3	**84.1**	**86.5**	**77.8**	**72.8**	**83.6**	**87.4**	**59.9**	**75.5**	**42.7**	**79.8**	**88.0**	**77.4**	**50.1**	**84.7**	76.0	**75.1**
	+ DF-FSOD	**13.9**	**26.3**	**58.5**	**44.9**	**8.4**	**30.4**	73.3	76.4	71.5	60.1	74.8	79.6	52.4	70.1	67.2	69.5	78.8	68.3	47.0	77.8	**77.5**	69.6
10	TFA [24] w/fc*	14.9	**59.3**	64.4	58.6	46.3	48.7	**88.4**	**87.5**	**79.2**	**73.4**	**86.7**	**88.7**	**64.9**	**83.4**	**67.5**	**82.2**	**88.3**	**86.4**	**56.1**	**87.9**	**77.4**	**79.9**
	+ DF-FSOD	**15.0**	57.5	**65.1**	**59.4**	**46.8**	**48.8**	77.0	79.5	71.4	66.5	75.5	78.9	54.6	71.9	66.8	70.6	79.0	75.1	48.0	79.9	66.1	70.7
	TFA [24] w/cos*	**14.2**	**60.7**	62.4	58.0	**46.9**	48.5	**88.9**	**86.3**	**78.3**	**72.9**	**85.5**	**88.5**	**61.7**	**83.3**	**69.7**	**81.5**	**87.7**	**86.1**	**54.4**	**86.9**	**78.2**	**79.3**
	+ DF-FSOD	13.4	60.0	**64.1**	**61.5**	45.9	**49.0**	79.8	81.3	71.5	65.3	74.3	78.9	55.3	72.7	66.1	74	78.9	77.3	47.8	80.1	69.3	71.5

TABLE III
MAP, AP@50 AND AP@75 IN THE NOVEL THIRD SET ON PASCAL VOC.

Shot	Method	Base			Novel		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
1	TFA [24] w/fc*	**46.5**	**75.7**	**49.9**	13.7	25.0	12.9
	+ DF-FSOD	43.0	68.8	47.1	**16.0**	**30.0**	**14.7**
	TFA [24] w/cos*	**46.8**	**75.1**	**49.8**	11.6	21.3	10.7
	+ DF-FSOD	43.1	69.6	46.9	**16.9**	**30.4**	**15.5**
10	TFA [24] w/fc*	**52.2**	**79.9**	**57.9**	**27.8**	48.7	**28.7**
	+ DF-FSOD	47.0	70.7	53.0	27.4	**48.8**	27.3
	TFA [24] w/cos*	**52.4**	**79.3**	**57.9**	**28.2**	48.5	**29.2**
	+ DF-FSOD	47.0	71.5	52.9	27.9	**49.0**	28.3

with other methods on the novel classes C_{novel}, we have to fix the shots of each class. Therefore, the experiments are implemented under the condition:

$$\forall c \in C_k, N_c = K \tag{1}$$

where $C_k = c_{base} \cup c_{novel}$. c_{base} and $c_{novel} \subset C_{novel}$ are base and novel classes of the experiment, respectively. N_c is the quantity of instances for each class c. Based on these constraints, we can increase or decrease the number of base classes C_{base} which only contain instances in base dataset D_{base}. We use the Eq. 1 to set up the training data of DF-FSOD shown in Fig. 2. In addition, we also impale the number of novel classes C_{novel} with the same benchmark novel dataset D_{novel} for fair comparisons, the results are shown in Tab. IV.

B. Two-stage fine-tuning

This subsection presents the two-stage approach to investigate the relation between the diversity categories and the recognizing the novel objects. Our framework is a two-stage detector Faster RCNN [8] using ResNet [3] as backbone. In the first phase, we train the entire model in the base dataset D_{base}. We then freeze almost the layers of the model during the training of the second phase. Only two last layers of the detector are fine-tuned in the novel dataset D_{novel}, as shown in Fig. 2.

Base training. In the first stage, we train the detectors in a large dataset containing the base classes C_{base} with the same loss function of Faster RCNN [8]. The ultimate goal of this stage is to gather helpful information from the source domain to transfer to the target domain as novel classes C_{novel}.

Novel fine-tuning. In the second stage, we differently create few-shot settings to investigate the affection of the number of categories. For **DF-FSOD**, we increase twice the number of base classes while holding novel classes C_{novel} as the benchmark. The training set contains K shots of each class in the traditional setup, including base and novel classes (i.e., 20 classes in PASCAL VOC).

On the other hand, based on the aforementioned hypothesis in Sec. III and the constraints Sec. IV-A, we design the experiments which are adopted from the TFA [24] setup. We assign randomly initialized weights to the new layers (the box predictor and classifier) and only fine-tune the two last layers with our experiments as shown in Fig. 2.

V. EXPERIMENTS AND DISCUSSION

In this section, we conduct the experiments to evaluate our approach in comparison with the previous methods on the existing FSOD benchmarks PASCAL VOC [10], [11]. The few-shot settings are based on [15], [16], [24] for fair comparisons. We also fix instances for each shot dataset to reduce the difference among the results because of choosing training data randomly. Therefore, we reimplement our baseline TFA [24] including two versions as TFA with fully connected (TFA w/fc) and TFA with cosine (TFA w/cos). Moreover, we provide various ablation studies and visualization in Sec. V-B.

A. Dataset and Settings

PASCAL VOC. Following the previous work [15], [16], [24], we evaluate all the baselines on FSOD challenge of PASCAL VOC 2007 [10] and 2012 [11]. The challenge splits 20 classes of VOC into 3 different sets in which 5 classes for novel set and the remaining 15 for base set, i.e., (“Bird”, “Bus”, “Cow”, “Mbike”, “Sofa”/ rest); (“Aero”, “Bottle”, “Cow”, “Horse”, “Sofa” / rest) and (“Boat”, “Cat”, “Mbike”, “Sheep”, “Sofa”/ rest). Each novel class contains K instances for novel phase ($K \in \{1, 2, 3, 5, 10\}$) and they have M_{base} = 15, M_{novel} = 5.
Implementation details. Our model adopts Faster RCNN [8] using ResNet-101 backbone [3] with Feature Pyramid Network [28]. We use SGD optimizer with an initial learning rate of 0.02, a mini-batch size of 8, momentum of 0.9 and the weight decay of 0.0001. The base model is trained with 36000 iterations, and the learning rate is divided by 10 at 24000 and 32000 iterations. In the fine-tuning stage, we train the model with two setups (the origin and the expanding). In the origin

Fig. 4. More illustration of our results on 1-shot dataset in novel third set of PASCAL VOC. The first row is the results of the baseline. The second row shows the results of our approach. In bad cases, our approach miss-classify some objects (the three last columns) but better both localize and classify the baseline on the others.

TABLE IV
THE ABLATION STUDY OF THE INCREASING NUMBER OF NOVEL CLASSES WHEN THE NUMBER OF BASE CLASSES IS 15. AP@50 AND MEAN OF NOVEL CLASSES IN THE NOVEL THIRD SET ON PASCAL VOC.

# shots	# cls	Mean	Boat	Cat	Mbike	Sheep	Sofa
1	16	11.6	11.6	-	-	-	-
	18	25.4	13.2	18.3	44.9	-	-
	20	21.3	13.0	19.3	43.1	23.4	7.6
2	16	14.4	14.4	-	-	-	-
	18	39.2	14.4	40.1	63.0	-	-
	20	38.2	14.2	36.8	59.4	37.4	43.2
3	16	10.3	10.3	-	-	-	-
	18	34.3	10.4	43.6	48.8	-	-
	20	40.6	10.6	40.6	47.8	58.4	45.3
5	16	18.4	18.4	-	-	-	-
	18	46.2	18.5	58.6	61.5	-	-
	20	48.0	16.2	57.9	60.6	60.0	45.5
10	16	14.5	14.5	-	-	-	-
	18	45.9	14.6	60.1	63.0	-	-
	20	48.5	14.2	60.7	62.4	58.0	46.9

setup, the models are trained with training data containing 20 classes for 7000 and 8000 iterations with learning rates of 0.004 and 0.0004, respectively. In the expanding setup, we randomly choose instances to expand the base classes to 30 subclasses. Therefore, the training data contains totally 35 classes consisting of 30 classes expanded from the base and 5 classes of the novel. In testing, we merge 30 expanded classes into 15 origin classes. The models in this setup are trained for 13000 and 14000 iterations with learning rates of 0.004 and 0.0004, respectively. In the experiments of ablation studies, we train the models with the same learning rate and linear scale iterations based on the number of classes with the origin setup.

B. Results and visualization

Main results. We first provide the average AP@50 of 3 novel sets on PASCAL VOC in Tab. I. Our approach outperforms the baselines in almost all experiments. DF-FSOD takes a large gap (4% AP@50) when labels are extremely scarce (1 or 2 shots) except for the second set. Our method achieves remarkable performance on all 3 novel sets compared to both TFA only fine-tuning FC layers and TFA applying the cosine similarity. It improves average $4.5-7.9\%$ on novel sets except for the second set. Using our setting, the models learn more discriminative information and achieve higher performance $(5-42\%)$ in the 1-shot dataset. These results demonstrate the potential of our approach, which is the first step for a new efficient approach in FSOD.

The proposed methods become less efficient with more training data, which probably occurs because of the native way to expand the subclasses in base classes. In the 1 or 2-shot dataset, the training data has high-discriminative information to support the learning process of models. However, due to randomly choosing the instances of expanding base classes, the signals in training data can confuse the models. Therefore, DF-FSOD yields minor improvement in 3, 5 and 10-shot datasets.

In Tab. II, we show the detailed results for each class, including both base and novel classes. Our approach improves the average performance of novel classes in 1 and 10-shot datasets compared with the baselines. As can be seen, the models with our approach better capture novel objects with scarce data than baseline. In the situation of the highly scarce available data as 1-shot, our DF-FSOD takes a significant improvement (from around 20 to 42%). In contrast, we have less efficiency (1% improvement) when having more labels (10 shots). In addition to AP@50, we show general evaluation in Tab. III, including mAP and AP@75 as well. However, they have a slight reduction in seen classes to trade off the high performance of unseen classes.

Visualization. Besides the above metrics, we also illustrate our results in Fig. 3 and Fig. 4. Our approach detects novel objects when maintaining the accuracy on the base classes. The good cases in Fig. 3 show that DF-FSOD better localizes and classifies both base and novel objects in the test set by utilizing more distinguished information from base classes. Even some

of our bad cases are shown in Fig. 4, our approach outperforms its baseline. Our model can better recognize small objects (the second and third columns) and localize occluded objects (three last columns).

Ablation studies. We conduct extensive experiments to show the affection of classes in FSOD. In Tab. IV, we fix M_{base} = 15 with $c_{base} = C_{base}$ of the novel third set and gradually increase the number of novel classes to get the observation. With $c_{novel} \subset$ {"Boat", "Cat", "Motorbike", "Sheep", "Sofa"}, mean AP@50 improves when increasing M_{novel} ($M_{novel} \in \{1, 3, 5\}$). However, the improvement of mean AP@50 is due to the high score of the additional novel classes.

The accuracy of the detector in class "Boat" only improves in the 1-shot dataset as soon as we add more training data (improvement 1.4% AP@50). In comparison, the performance of "Boat" is kept or even reduced when adding more novel classes (16, 20 classes). This phenomenon happens because of the ambiguity of information from additional classes in 2, 3, 5, 10-shot datasets, making the models create ineffectively distinguished features. Expanding or increasing the number of classes, therefore, is essential to improve the performance of unseen classes with scarce training data. However, the expanding classes have to contain highly discriminative information to support the detectors.

VI. Conclusion

In this paper, we present a comprehensive investigation about the affection of the number of classes in FSOD. Studies show that expanding or increasing the number of classes is essential to improve the performance of unseen classes with scarce training data. However, the added classes must contain highly discriminative information to support the models. We also present a novel approach (DF-FSOD) to improve the detectors in FSOD to capture novel objects better. Our approach is to create models capable of generating highly distinguished features, which helps the model increase the ability to detect objects, including base and novel classes, effectively.

Acknowledgments. This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2021-20.

References

[1] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019.

[2] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6Tm1mposlrM

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[5] B. Artacho and A. Savakis, "Unipose: Unified human pose estimation in single images and videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[6] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[7] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, 2016.

[9] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, 2010.

[11] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, 2015.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[13] Y. Wang, Z. Xu, H. Shen, B. Cheng, and L. Yang, "Centermask: Single shot instance segmentation with point representation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[14] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[15] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *International Conference on Computer Vision*, 2019.

[16] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *International Conference on Computer Vision*, 2019.

[17] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein, "Repmet: Representative-based metric learning for classification and few-shot object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[18] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *International Conference on Computer Vision*, 2017.

[20] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[21] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016.

[24] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *International Conference on Machine Learning (ICML)*, 2020.

[25] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, 2013.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, 2015.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.