Few-shot Object Detection via Improved Classification Features

Xinyu Jiang[1], Zhengjia Li[1], Maoqing Tian[2], Jianbo Liu[3], Shuai Yi[2], Duoqian Miao[1*]
[1] Tongji University [2] Sensetime Research [3] Chinese University of Hong Kong
{xinyujiang, zjli1997, dqmiao}@tongji.edu.cn {tianmaoqing, yishuai}@sensetime.com liujianbo@link.cuhk.edu.hk

Abstract

Few-shot object detection (FSOD) aims to transfer knowledge from base classes to novel classes, which receives widespread attention recently. The performance of current techniques is, however, limited by the poor classification ability and the improper features in the detection head. To circumvent this issue, we propose a Multi-level Feature Enhancement (MFE) model to improve the feature for classification from three different perspectives, including the spatial level, the task level and the regularization level. First, we revise the classifier's input feature at the spatial level by using information from the regression head. Secondly, we separate the RoI-Align feature into two different feature distributions in order to improve features at the task level. Finally, taking into account the overfitting problem in FSOD, we design a simple but efficient regularization enhancement module to sample features into various distributions and enhance the regularization ability of classification. Extensive experiments show that our method achieves competitive results on PASCAL VOC datasets, and exceeds current state-of-the-art methods in all shot settings on challenging MS-COCO datasets.

1. Introduction

Recently general object detection achieves great improvement. Numerous innovative techniques [17, 4] have been proposed, but most of them can not deliver fairly satisfactory performance in the few-shot setting. Therefore, there is a clear difference in intelligence between humans and these manual algorithms, as humans can recognize new objects even after only a few exposures. Learning from a small number of instances is significant for object detection in realistic scenes.

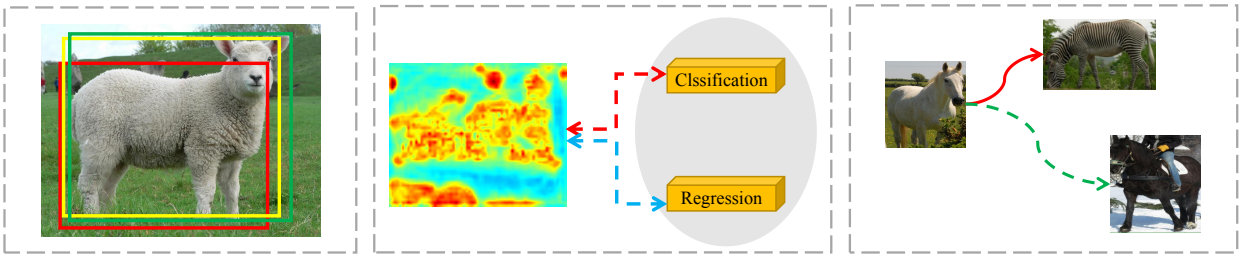
Few-shot object detection (FSOD) is a challenging task that combines few-shot learning and object detection. Given the base classes with abundant training data and the novel classes with few annotations, FSOD trains a model which learns general knowledge from the base classes and then leverages them on the novel classes. Previous models can be divided into two classes: *meta-learning* based and *transfer learning* based methods. The meta-learning based methods aim to solve the FSOD task in meta-learning paradigm, which mainly follows the uniform/adaptive sampling scheme to generate tasks at each episode[12, 34, 30]. For the transfer learning based methods, they first train a model on base classes with numerous instances, and then fine-tune this model on novel classes with only a few instances.

One widely held belief has emerged with the development of few-shot object detection, which is that the classifier's performance is the main bottleneck for this task. Some works have been proposed recently to improve the classification performance in FSOD[25, 13]. Despite focusing on addressing the classification problem for few-shot object detection, these new methods ignore that the input of the classifier was based on the original proposal features. The following issues could arise if original features are used:

Spatial shift. In the widely used Faster R-CNN architecture, the regressor and classifier use the same feature from the stem representation of the RoI-Align feature to perform regression and classification simultaneously. In the general object detection task, there are many annotated examples used in the training stage and the model is robust for the diversity between the proposal feature and accurate bounding box feature. However, low-quality proposal features limit classification performance in few-shot settings. The classifier's performance is irreparably harmed by the discrepancy between accurate bounding box features and proposal features. As shown in Figure 1a, the features used in classification are inaccurate, which causes poorer classifier performance in the few-shot setting.

Task conflict. The goals of classification and regression are not the same because they are two distinct tasks. Regressor focuses on locating objects meanwhile classifier tries to distinguish different categories. Therefore, Faster R-CNN architecture suffers from the conflicting objectives

*Corresponding author.



(a) (b) (c)

Figure 1: Potential issues in Few-shot Object Detection. (a) denotes spatial shift issue. Green bounding box: Correct coordinate of the object. Red bounding box: Object coordinate generated by RPN. Yellow bounding box: Coordinate generated by R-CNN based on the proposal. (b) denotes task-level issue. The same proposal features are difficult to represent classification and regression tasks at the same time, and there may exist conflicts between tasks in the backpropagation stage. (c) denotes overfitting issue. In the few-shot setting, it isn't easy to pinpoint the feature distribution of the horse so the horse is misidentifying as the zebra.

of classification and regression during training, and features generated by the RoI-Align operation cannot consider this conflict. The ability of the model to locate and classify simultaneously in object detection is generally enabled by sufficient training data, which mitigates this phenomenon. However, in the few-shot setting, the struggle between location and classification is more intense.

Severely overfitting. Few-shot object detection expects to learn all representations of a class with very little training data, which is difficult for general classifiers. Particularly in the case of few-shot settings, it is very simple for the model to overfit the training data, and it is challenging to discriminate objects in the feature space that are far from the training data. As a result, the model has trouble differentiating between classes like horse and zebra which are similar.

In this paper, we focus on improving the quality of classification features in few-shot object detection model. With few-shot setting, it is difficult for features to be greatly enhanced by a single level, so we propose a Multi-Level Feature Enhancement (MFE) method to enhance classification features on three specific levels, including spatial-level, task-level, and regularization-level, in light of the phenomenon we described above. By altering the spatial locations of the classification features, the spatial-level module enhances the features of classification. By focusing on different channels of features, the task-level module decouples the tasks of localization and classification. The regularization-level module improves the classifier's regularization capabilities, which resolves the issue of inconsistency between training and inference in few-shot setting. Through the fusion of three feature enhancement modules, MFE enhances the original proposal features into features adapted for classification. The experimental results show that MFE greatly enhances the performance of the two-stage few-shot detector.

The main contributions of our approaches are three-fold:

- We point out the existing problems in few-shot object detection in the view of imperfect features for the classifier, which are not crucial in general object detection.
- We propose the Multi-level Feature Enhancement(MFE) to improve detection features from spatial, task, and regularization levels.
- Our approach achieves competitive results on COCO and PASCAL VOC benchmark, which demonstrate the effectiveness of our framework.

2. Related Work

Few-shot classification. As a challenging and meaningful problem, many methods [8, 28] have been proposed for the few-shot classification to improve the quick adaption ability from base classes to novel classes with only a few samples. These methods attempt to learn many training tasks to solve a new unseen few-shot task. In each iteration, the model learns from a specific n-way k-shot task to leverage task-level meta knowledge, which is known as meta-learning. Some approaches in meta-learning optimize the gradient descent procedure to find a good initialization for new task[8, 14]. Some other approaches aim to learn a better embedding space for few-shot learning[28, 24, 26, 3]. In addition to meta-learning paradigm, there are also some methods that explore the fine-tune paradigm in few-shot learning[18, 3, 27]. Chen *et al.* [3] proposed that a simple pretrain and fine-tune model with the last classifier layer can also get a competitive performance compared to meta-learning.

Few-shot Object Detection. There are two streams of preliminary work for the FSOD task. One of the most general approaches is the meta-learning based approach. Meta-learning algorithms abstract the training process into a task

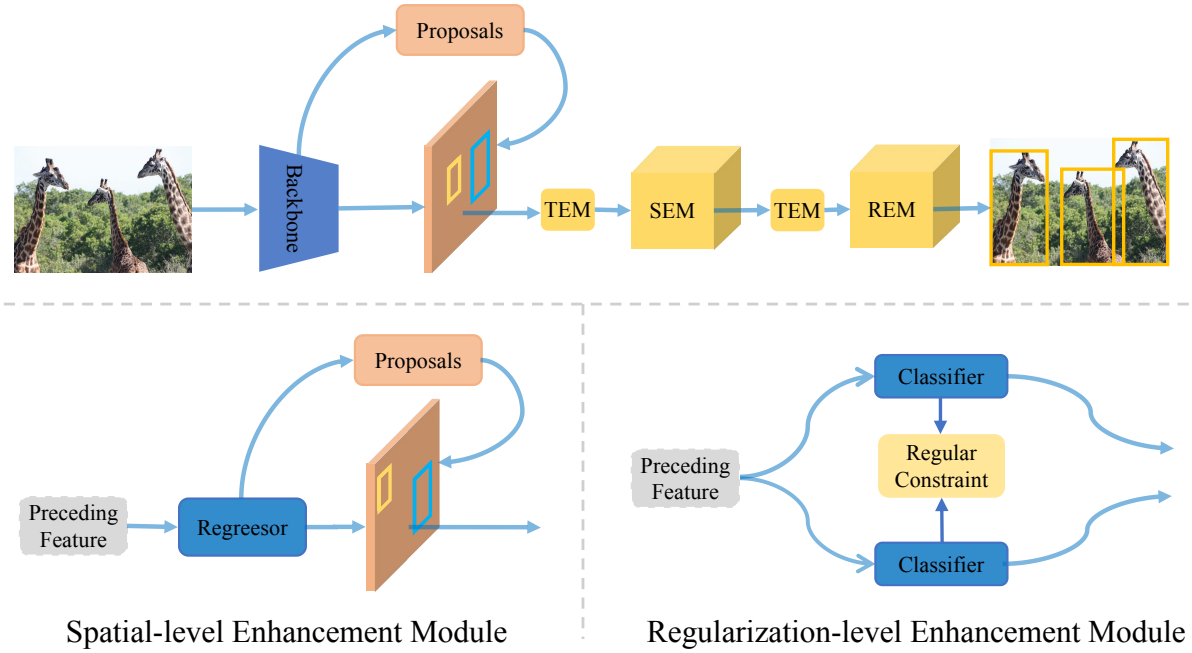


Figure 2: MFE architecture. Instead of standard R-CNN architecture, We enhance the features for the classifier from spatial, task, and regularization levels. SEM, TEM, and REM denote spatial-level, task-level, and regularization-level enhancement modules respectively.

paradigm, utilizing a set of support examples to predict the objects in query images. Kang *et al.*[12] applied meta-learning firstly in FSOD task combining YOLO architecture via feature reweighting[21]. Yan *et al.* [34] proposed to meta-learn an RoI module of Faster R-CNN architecture. Xiao *et al.* [33] defined a simple yet effective unifying framework that tackles both few-shot object detection and few-shot viewpoint estimation. Fan *.et al.* [6] proposed a general few-shot object detection network that learns the matching metric between image pairs. Wu *et al.* [31] enhanced object features using a universal prototype in meta-learning way.

Another stream is the transfer learning based approach. Chen *et al.*[1] involved this problem in a transfer learning way by combining SSD[16] and Faster R-CNN fashion [22]. Wang *et al.* [29] pointed out that fine-tune only the last layer of the existing detector is crucial to the FSOD task. Wu *et al.* [32] proposed a multi-scale positive sample refinement to enrich object scales. Recently, more and more works focus on improving classification performance. Sun *et al.*[25] applied the contrasted loss to distinguish similar class meanwhile Li *et al.*[13] leveraged a class margin loss technique to balance inter and intra class margins. Qiao *et al.*[20] captured the conflicts between RPN and R-CNN module and introduced a decouple module to solve this problem.

3. Methods

In this paper, we propose a Multi-Level Feature Enhancement (MFE) method to improve classification features from three particular perspectives, which is illustrated in Figure 2, including the spatial-level enhancement module (SEM), the task-level enhancement module (TEM) and the regularization-level enhancement module (REM). We introduce the problem definition in Section 3.1 and the proposed three modules of the different levels in MFE in Section 3.2, 3.3, 3.4.

3.1. Problem Definition

We first give a formal few-shot object detection definition followed by Kang[12]. Given two sets of data C_{base} and C_{novel}, C_{base} denotes the abundant annotated instances in base classes, and C_{novel} denotes the few annotated instances in novel classes. The intersection categories between C_{base} and C_{novel} are $\emptyset$. We aim to obtain a few-shot object detection model by sufficiently exploiting generalized knowledge from base classes and transferring them into novel classes. Usually, there are only k instances in C_{novel} per class, mentioned as k-shot object detection.

We mainly employ transfer learning approaches for our model training and testing, the same as previous transfer learning based approaches[29, 20]. Model training can be summarized as two stages: in the first stage, our model is trained on base classes and learns generalized knowledge

Stage/IoU	0.50:0.90	0.90:1.00
base training	39.89	0.34
10-shot training	24.61	0.07

Table 1: Comparison of the number of proposals generated by RPN in the two training phases per image on MS COCO benchmark.

such as foreground bounding box features and basic information to distinguish objects. In the second stage, also called the novel training stage, to utilize the knowledge learning from the first stage, we further fine-tune this model based on novel categories. In the testing stage, our model aims to detect objects belonging to novel categories.

3.2. Spatial-level Enhancement

For few-shot object detection, we observe that there is a general problem: The proposal features utilized in classification are not reliable, as illustrated in Figure 1a, which is the major cause for inferior classifier results in the few-shot setting. To demonstrate this opinion, we counted the number of proposals generated by RPN in the two training phases of FSOD in Table 1. Although more low-quality positive sample proposals were generated in the base training phase than in the 10-shot training phase, the difference was not significant. However, for high-quality proposals that have IoU greater than 0.9 with gt, there is a nearly five-fold difference in the number of proposals generated in the two phases. As a result, compared to the base training stage, the high-quality proposals are much less in 10-shot training stage. The low-quality proposal features damage the performance of the classifier in the detection head irreversibly.

To solve the above problem, we exploit R-CNN as a stronger RPN to provide more accurate candidate bounding boxes, and the Spatial Enhancement Module(SEM) is proposed to improve the final classification results, which is illustrated in Figure 2. We first modify the regression module of R-CNN in a class-agnostic manner before incorporating the SEM model. Our regression module in MFE only recognizes the foreground object and does not produce a bounding box for each category. Although this setting slightly lowers performance, it helps us recognize various categories and better adapt to our spatial module. Then, to provide a stronger RPN for the classification task, we sequentially connect the modules for regression and classification. The classification module performs better with such a framework in place.

Beyond our SEM, we also discuss variants of SEM. One simple thought is that we can also get classification results from the original proposal as additional supervision. Similarly, another regression loss can be calculated by the updated bounding boxes feature. We refer to these two variants of SEM as SEM-c and SEM-r. Experimental analysis in Section 4.4 demonstrates that these two auxiliary heads are not necessary and SEM outperforms these two variants.

In section 4.5, we analyze the performance impact of SEM between sufficient data setting and few-shot setting to demonstrate that spatial shift is more severe in FSOD.

3.3. Task-level Enhancement

In standard R-CNN, classifier and regressor employ the same feature to classify and locate. Classification and regression, however, serve different purposes. While the classifier tries to distinguish between various categories, the regressor focuses on locating the boundary of objects. During training, they suffer from the conflict between the classification and regression objectives, and the features generated by RoI-Align cannot take this conflict into account.

We created a task enhancement module (TEM) to decompose the features in the original space into a unique space for each task to address the aforementioned problems. Our primary goal is to more effectively address the classification and regression tasks, avoiding conflicts between these two distinct tasks. As shown in Eq 1, we employ a channel-wise attention mechanism to achieve this.

$$\begin{aligned} D(z, \theta, \phi) &= z \otimes s \\ s &= \sigma(W_2(ReLU(W_1(Pool(z)), \phi)), \theta) \end{aligned} \tag{1}$$

As shown in Eq 1, during the forward propagation, the convolutional features of each proposal $z \in \mathbb{R}^{b \times c \times 4 \times 4}$ are collapsed into a vector by max pooling, losing spatial information. Then these features are transformed by a linear layer W_1 into a smaller feature space $\mathbb{R}^{b \times c//16 \times 1 \times 1}$. After that, a $ReLU$ layer and another linear layer W_2 to ascend dimension of feature into $\mathbb{R}^{b \times c \times 1 \times 1}$ and calculate the attention score $D(z)$ through sigmoid function σ. ϕ and θ denote the parameters in W_1 and W_2 respectively. As a result, we get an attention score s in the direction of image height and width. Then we output a new representation for the RoI feature by combing channel-wise attention and origin feature. The $\otimes$ denotes element-wise multiplication, σ denotes the sigmoid activation function.

Are there better transform layers?

We explore several kinds of modules to achieve our TEM, including linear transform, spatial-wise attention and channel-wise attention. The experiment results show that channel-wise attention outperforms other designs by a large margin. Furthermore, We attempt to use channel-wise attention without the task-specific adaptor, and the results show that the idea of task-specific adaptor is the primary element that can improve performance. The detail of the experiments is discussed in Section 4.4.

Method/Shots	1-shot	2-shot	3-shot	5-shot	10-shot	30-shot
FRCN-ft[30]	1.0*	1.8*	2.8*	4.0*	6.5	11.1
TFA[29]	4.4	5.4	6.0	7.7	10.0	13.7
MPSR[32]	5.1	6.7	7.4	8.7	9.8	14.1
FSDetView[33]	4.5	6.6	7.2	10.7	12.5	14.7
Meta Faster R-CNN[9]	5.1	7.6	9.8	10.8	12.7	16.6
CME[13]	-	-	-	-	15.1	16.9
FCT[10]	5.6	7.9	11.1	14.0	17.1	21.4
DeFRCN[20]	9.3	12.9	14.8	16.1	18.5	22.6
DAnA-FasterRCNN[2]	-	-	-	-	18.6	21.6
MFE(ours)	**10.5**	**13.5**	**15.8**	**17.9**	**20.1**	**24.1**

Table 2: Few-shot detection performance(mAP) on MS-COCO dataset. We evaluate 1,2,3,5,10 and 30 shot performance over multiple runs. The bold font represents the best result. '-' indicates no reported results.

3.4. Regularization-level Enhancement

The fundamental cause of few-shot object detection difficulty is a serious shortage of data. When only one instance of a class has been observed, the detector is unable to acquire the necessary information about features to distinguish between related classes.

To reduce the overfitting problem, we expect to enhance the regularization ability of the model and guide the model to identify classes based on part of the information of features. To achieve this, we design a simple but efficient regularized feature enhancement module with a Regularized Consistent (RC) Loss as follows:

$$\mathcal{L}_{RC} = \mathcal{F}(g_1, g_2) \tag{2}$$

where g_1 and g_2 are different samples of the same feature. $\mathcal{F}$ is a measure function of whether g_1 and g_2 are consistent. We sample g_1 and g_2 using dropout technique and we use Kullback-Leibler (KL) divergence to implement $\mathcal{F}$. Therefore, the total loss of MFE can be summarized as:

$$\mathcal{L}_{MFE} = \mathcal{L}_{SEM} + \mathcal{L}_{RPN} + \alpha[\mathcal{L}_{RC_1} + \mathcal{L}_{RC_2}] \tag{3}$$

$\mathcal{L}_{SEM}$ includes a regression loss and two classification loss for g_1 and g_2. α is the hyperparameter that controls the weight of RC loss. By adding such regularization-level enhancement, our model augment the data from the feature level, while being robust to the different distribution of data. The method can be easily applied to different model structures, while more complex sampling methods can also be considered.

4. Experiments

In this section, we first introduce more implementation details and extensive experiment results. Then we give additional ablation studies and visualizations to evidence the effectiveness of our work.

Implementation Details We use Faster R-CNN as our detection model and choose standard ResNet-101 [11] pretrained on ImageNet[23] as the backbone. We re-implement DeFRCN[20] based on detectron2 as the baseline. Specifically, we modify the regression head into a class-agnostic fashion. Both the base training stage and fine-tune stage adopt SGD optimizer with a mini-batch size of 16 and the batch size in proposal sampling is 512. The initial learning rates are 0.02 and 0.01 for base training and fine-tune training stage respectively. α in RC Loss is 1. We observed that the model on MS COCO requires more iteration to convergence due to more categories compared to PASCAL VOC, so the model is trained for 110000 iterations in COCO and 20000 iterations in PASCAL VOC in total.

4.1. Experiment Benchmark

MS COCO. MS COCO[15] is a challenging benchmark in object detection, especially in few-shot setting. Following previous works[12, 29], the 80 categories are divided into 60 base categories and 20 novel categories. All training data come from MS COCO 2014 trainval dataset and 5K images from minival dataset are used as testing data. K-shot of novel instances are randomly sampled from unseen novel classes, and here the k = 1,2,3,5,10 and 30 by convention. We evaluate our model on the COCO-style *mAP*.

PASCAL VOC. PASCAL VOC 07+12 dataset[5] consists of 20 categories. Following existing works[12, 29],

Method/Shots	Novel Set 1				Novel Set 2				Novel Set 3			
	1	2	3	10	1	2	3	10	1	2	3	10
Meta R-CNN[34]	19.9	25.5	35.0	51.5	10.4	19.4	29.6	45.4	14.3	18.2	27.5	48.1
TFA [29]	39.8	36.1	44.7	56.0	23.5	26.9	34.1	39.1	30.8	34.8	42.8	49.8
MPSR[32]	41.7	42.5	51.4	61.8	24.4	29.3	39.2	47.8	35.6	41.8	42.3	49.7
CME[13]	41.5	47.5	50.4	60.9	27.2	30.2	41.4	46.8	34.3	39.6	45.1	51.5
FSCE[25]	44.2	43.8	51.4	**63.4**	27.3	29.5	43.5	**50.2**	37.2	41.9	47.5	**58.5**
SRR-FSD[35]	47.8	50.5	51.3	56.8	32.5	35.3	39.1	43.8	40.1	41.5	44.3	46.4
DeFRCN[20]	53.6	**57.5**	**61.5**	60.8	30.1	38.1	**47.0**	47.9	48.4	**50.9**	**52.3**	57.4
MFE (ours)	**55.0**	55.5	59.2	59.7	**34.7**	**38.2**	44.1	46.4	**49.5**	44.2	47.3	55.4

Table 3: Experimental results on VOC dataset . We use AP_{50} as metrics and the evaluation performed over 3 different splits.

Method/Shots	mAP				AP_{75}			
	1	2	3	10	1	2	3	10
Baseline	32.9	34.7	37.4	38.1	34.5	37.3	39.0	41.9
MFE (ours)	**33.1(+0.2)**	**35.5(+0.8)**	**38.5(+1.1)**	**41.0(+2.9)**	**35.2(+0.7)**	**38.7(+1.4)**	**42.4(+3.4)**	**44.3(+2.4)**

Table 4: mAP and AP_{75} results on VOC dataset over Novel Set 1.

SEM	TEM	REM	5-shot	10-shot	30-shot
			15.7	18.5	22.6
✓			16.8	19.2	22.4
	✓		17.3	19.2	22.8
		✓	16.5	18.8	22.7
✓	✓		17.5	19.8	24.1
✓		✓	17.6	19.6	22.7
	✓	✓	17.5	19.3	23.1
✓	✓	✓	17.9	20.1	24.1

Table 5: Ablation studies about MFE performance on MS COCO.

Method/Shots	10-shot	30-shot
FsDetView[33]	6.7	10.0
ONCE[19]	13.7	-
MPSR[32]	15.3	17.1
FRCN-ft[30]	18.1	18.6
TFA[29]	27.9	29.7
Retentive R-CNN[7]	**32.1**	**32.9**
MFE (ours)	31.6	**32.9**

Table 6: Overall class results on COCO under 10,30-shot setting.

there are three random splits used for few-shot object detection, referred to as novel split 1,2, and 3. Each split includes 15 base classes and 5 novel classes. All base categories come from PASCAL VOC 07+12 trainval sets and we report AP_{50} for novel classes on PASCAL VOC test set.

4.2. Comparison Results

MS COCO. Table 2 shows our main evaluation results on the challenge COCO benchmark. Our model has an inspirational improvement compared to previous works. Compared to previous SOTA work[20, 2, 10, 13], our MFE outperforms them in all setups, by 1.2%,0.6%,1.0%,1.8%,1.5%,1.5% in terms of mAP on 1,2,3,5,10 and 30 shot respectively. To the best of our knowledge, we are the first to achieve 10% in 1-shot setting. What's more, our model almost halves the iterations that achieve convergence in fine-tune stage compared to our baseline. Furthermore, MFE also has the ability to detect in the Generalized FSOD(G-FSOD) setting, which we will discuss in Section 4.3.

PASCAL VOC. We present VOC evaluation results in Table 3, on three common splits. Our model gets competitive results, and in 1-shot setting our model ranks best in all splits, demonstrating our model addresses fewer-shot setting problems better. To illustrate the effectiveness of MFE even further, we evaluated the effectiveness of baseline and MFE under mAP and AP_{75} metrics in Table 4. Under the stricter positive sample determination, the MFE increased significantly from the baseline.

Method/Shots	1-shot	2-shot	3-shot
SEM-r	6.4	9.7	11.9
SEM-c	8.6	10.3	12.4
SEM	9.6	12.8	15.7

Table 7: Comparison results about our SEM and other similiar architecture.

Architecture/Shots	10-shot	30-shot
Linear Layer	18.1	22.1
Spatial-Attn	19.4	23.1
Channel-Attn(Avg Pool)	19.9	23.5
Channel-Attn(Max Pool)	19.9	24.1
Channel-Attn with shared weights	19.8	22.8

Table 8: Comparison results about channel-wise attention in TEM and other architecture.

4.3. Generalized Few-Shot Object Detection

Generalized few-shot object detection (GFSOD) not only pays attention to the performance in novel classes but also concerns the overall categories performance of the few-shot object detection methods. In GFSOD, k-shot of instances of each base category is also involved in fine-tune stage. It evaluates the incremental learning ability of few-shot learning model without forgetting. We report overall class results on COCO in Table 6. The competitive results compared to SOTA demonstrates that MFE can learn without forgetting.

4.4. Ablation

Components of our proposed MFE. We conduct the ablation study of the MFE module in Table 5. Compared to baseline, in few-shot setting, SEM improves the baseline's performance in every shot and brings mAP improvement up to 1.1% in 5-shot setting, which is huge progress considering the difficulty of this task. Additionally, all three of our modules help enhance the performance of detection. They can be thought of as cooperating orthogonally because they work on different dimensions. The three modules work together to obtain the best results.

The difference between other similar architecture and SEM. As discussed in Section 3.2, we also design two variants of SEM, referred to as SEM-r and SEM-c respectively, to demonstrate that our SEM is efficient compared to similar architecture. They both use sequential regression and classification in R-CNN but they generate another regression and classification module to assist optimization. As shown in Table 7, SEM-r performs far away from our SEM due to excessive concentration on regression is not beneficial for classification, which is the most significant task in FSOD as we mentioned before. SEM-c also decreases performance due to features from original proposals are not accurate, which is also the motivation of our SEM module.

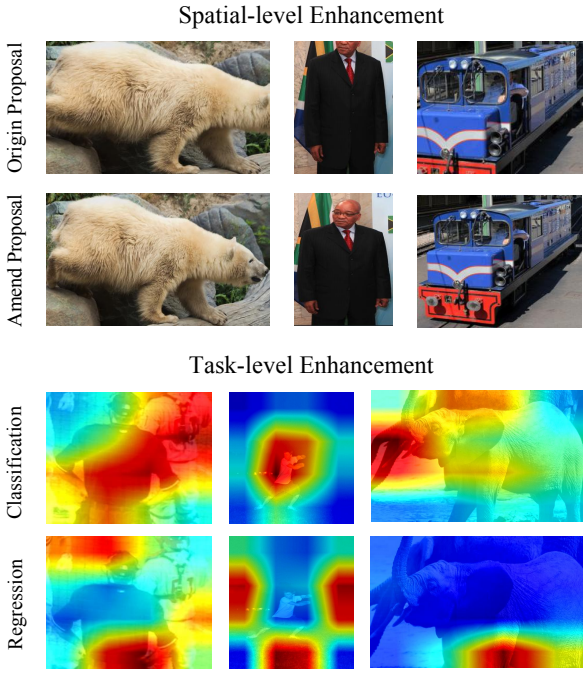


Figure 3: The visualization results of each of the spatial and task levels enhancement.

The difference between TEM and other similar architecture. We design our TEM using channel-wise attention and during the experiment, we explore the different types of transform layers and compare the performance between them in Table 8. In conclusion, we find that channel-wise attention outperforms spatial-wise attention and linear transformation layer by a large margin. Meanwhile, we also conduct experiments on the different forms of pooling in channel-wise attention, as summarized in the 3-4 rows. Max Pooling is a better option as a result.

As shown in the last row of Table 8, we also report the performance of TEM with shared weights, *i.e.*, classification and regression modules use the same feature from the attention layer, indicating that the feature is not disentangled. The results demonstrate that the improvement brought by TEM comes from both attention mechanism and feature disentanglement. Additionally, there is a 0.6 percent average difference between them in 1, 2, 3, and 5 shots.

4.5. Analyse and Visualization

There is a significant performance difference between general and few-shot object detection, and based on our in-

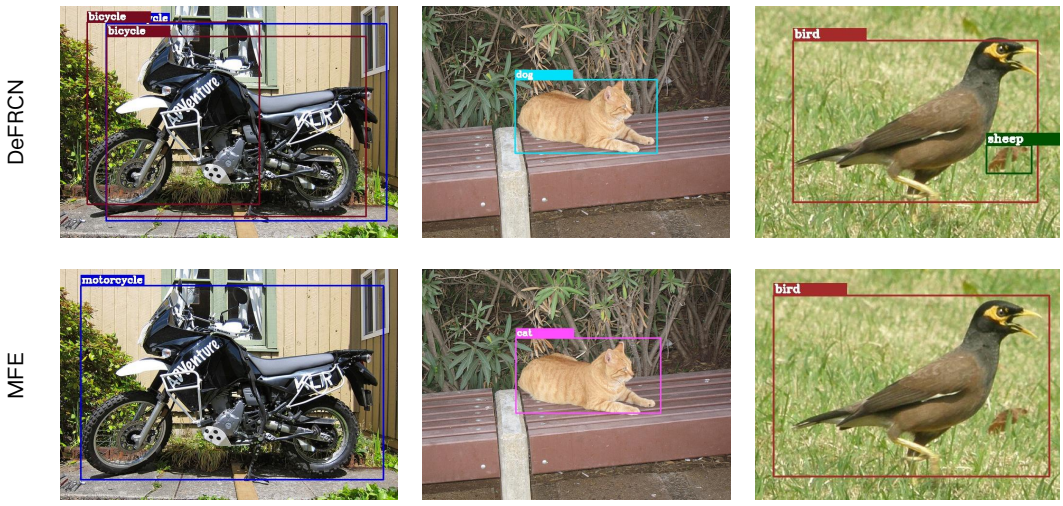


Figure 4: Visualization results of bad cases rescued by MFE.

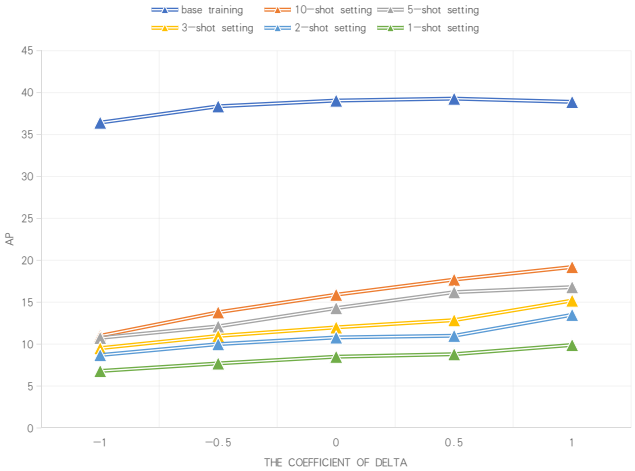


Figure 5: The influence of offset between few-shot setting with sufficient data setting.

vestigation, we verify that model is more sensitive to features under few-shot setting. As shown in Figure 5, delta denotes regression offset and we change the coefficient of delta to observe the performance of the detector. Denoting the coefficient of the delta by x, $x = 0$ means the original proposal feature, $x = 1$ means the proposal feature is updated in full dependence on the regression results, and $x < 0$ means the proposal is updated in the opposite direction of the offset. While the delta does not significantly affect the mAP result during the base training stage, it does in the few-shot setting. It can be explained that few-shot detector is more sensitive to spatial shift. There are only few features so spatial jitter influences the detector observably.

Visualization We visualize the improved features used in MFE in Figure 3. After spatial-level enhancement, origin proposals have been rectified obviously. In the task-level enhancement module, proposal features have been transformed into different feature maps to adapt to classification and regression tasks respectively.

We also provide qualitative results of MFE and DeFRCN in Figure 4. Under the challenge 10-shot MS COCO setting, we visualize the bounding boxes whose confidence scores are greater than 0.5 in both methods. Better performance of classification can be observed in these images, especially in some confused categories, which demonstrates our consideration in the view of feature is effective.

5. Conclusion

In this work, we propose a novel architecture for few-shot object detection in the view of features. We point out that the quality of features used in classification is significant for FSOD and involves a novel architecture referred to as MFE to improve it from three orthometric perspectives. With the sequential design of R-CNN, deployment of two task-specific adaptors, and a regularization consistent module, MFE enhances the performance of the classifier principally in spatial, task, and regularization levels respectively. In PASCAL VOC and MS COCO benchmarks, our model achieves competitive results, especially in MS COCO, we achieve the best performance in every setting.

Acknowledgement

This paper is partially supported by the National Natural Science Foundation of China (Serial No.61976158 , No.62163016, No.62006172, No.61976160, and No.62076182), and the Jiangxi "Double Thousand Plan", and the Jiangxi Provincial natural science fund (No. 20212ACB202001).

References

[1] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[2] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, Wen-Chin Chen, and Winston H Hsu. Dual-awareness attention for few-shot object detection. *arXiv preprint arXiv:2102.12152*, 2021.

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[4] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.

[7] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4527–4536, 2021.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[9] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022.

[10] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5321–5330, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[13] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2021.

[14] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[15] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[17] Z Liu, Y Lin, Y Cao, H Hu, Y Wei, Z Zhang, S Lin, and B Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arxiv 2021. *arXiv preprint arXiv:2103.14030*.

[18] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[19] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020.

[20] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021.

[21] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[24] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[25] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021.

[26] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[27] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer*

Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 266–282. Springer, 2020.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.

[29] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

[30] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9925–9934, 2019.

[31] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9567–9576, 2021.

[32] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020.

[33] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision*, pages 192–210. Springer, 2020.

[34] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9586, 2019.

[35] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-stable few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8782–8791, 2021.