Feature Synthesis for Few-Shot Object Detection

Chenchen Tao[1], Song Chen*[2], Yi Chen[1], Xiaojie Cai[1], and Chong Wang[1]

[1] Ningbo University, Zhejiang 315211, China,
[2] H3C Company, Zhejiang 310000, China,
2011082274@nbu.edu.cn

Abstract. The Few-Shot Object Detection (FSOD) task aims to detect novel instances in scenarios with limited data. Nevertheless, the feature distribution of the novel class can be easily influenced by the distribution of features from the base classes. This paper introduces the Feature Synthesis for Few-Shot Object Detection algorithm, leveraging Generative Adversarial Networks to generate visual features for novel classes. By combining semantic embeddings with real visual features, the generator is trained to enhance the correlation between synthetic features and their corresponding categories. Class prototypes are computed based on real features, and contrastive loss guides the constraint of synthetic feature distribution, improving model performance. Additionally, the algorithm incorporates Pseudo Margin Evaluation loss to calculate instance uncertainty scores and increase discrimination power. Experimental results on the MS-COCO dataset demonstrate the algorithm's effectiveness with significant performance gains.

Keywords: few-shot object detection, feature synthesis, contrastive learning

1 Introduction

In the domain of few-shot object detection (FSOD), two primary research directions have emerged. The first direction is meta-learning-based approaches, which use a stage-wise and periodic meta-training paradigm to train a meta-learner capable of transferring knowledge from base classes. Meta R-CNN [1] introduces meta-learning for channel-wise attention layer adaptation in the RoI head, improving object detection performance. Meta-DETR [2] leverages a meta-learning strategy to exploit inter-class correlations and effectively utilize correlations among different classes. Other methods such as FSIW [3] and TFA [4] improve upon the meta-learning paradigm by employing more complex feature aggregation techniques and introducing balanced datasets. FSOD-UP [5] focuses on learning invariant object features from all categories and enhancing them using a consistency loss. SRR-FSD [6] proposes a semantic space constructed from word embeddings, training the detector to project target instances onto this semantic space.

In this paper, we address the challenges of few-shot learning and few-shot object detection by proposing novel approaches that leverage meta-learning and

transfer learning paradigms. Our aim is to enhance the model's ability to recognize and classify novel classes with limited labeled examples. To achieve this, we introduce a Feature Synthesis framework for Few-Shot Object Detection, which exploits semantic information and synthesizes diverse visual features using generative adversarial networks (GANs) [7].

Our work stands out as the first attempt to incorporate feature synthesis into the few-shot object detection (FSOD) task. By combining semantic information, we propose an algorithm that effectively addresses the scarcity of novel class samples in FSOD. This algorithm constrains the synthetic features by leveraging class prototypes, ensuring that the generated visual features closely resemble real features in terms of distribution.

Moreover, we tackle the challenge of uncertainty in synthesized features by employing the Pseudo Margin Evaluation (PME) loss. This loss function enables us to exploit the uncertainty associated with synthesized features, making even low-quality features useful for the FSOD model.

Through comprehensive experiments conducted on benchmark datasets, we demonstrate the effectiveness and superiority of our proposed methods in the field of few-shot object detection. Our contributions include addressing data scarcity, enhancing feature synthesis with semantic information, and effectively utilizing the uncertainty of synthesized features. These advancements significantly improve the generalization performance of few-shot object detection models.

The rest of this paper is organized as follows. The approach of our feature synthesis algorithm is presented in Section 2. The experimental results and analysis are given in Section 3. We conclude this paper in Section 4.

2 Method

2.1 Problem Definition and Framework Overview

Let D_B be the dataset containing base class object images and D_N denote the dataset with novel class object images. The labels for base and novel classes are $Y_B = \{1, ..., B\}$ and $Y_N = \{B + 1, ..., B + N\}$, respectively, where B and N represent the total number of base and novel classes, and $Y_B \cap Y_N = \emptyset$.

The framework begins by extracting features from input images $x \in D_B$ using ResNet-101 [8] and FPN [9] as the backbone network. The Region Proposal Network (RPN) generates candidate proposals, which are then processed through RoI pooling to obtain fixed 1024-dimensional features. The Faster R-CNN [10] is trained using D_N with k instances per class, while freezing the backbone network's parameters to prevent overfitting. This results in a standard FSOD model F_n, used to extract visual features $f_B \in R^{1024}$ and $f_N \in R^{1024}$ for base and novel classes. To utilize novel class knowledge, semantic embeddings from the CLIP [11] model are introduced as $A = \{a_0\} \cup A_B \cup A_N$, where a_0 represents background class embeddings and A_B and A_N are sets of base and novel class semantic embeddings, respectively. The overall framework of the proposed based

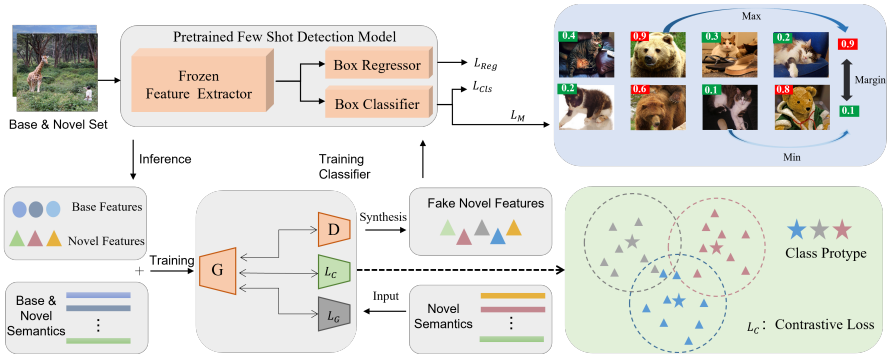

Fig. 1: Illustration of the proposed overall framework. Our framework is mainly divided into a feature synthesis module, a contrastive loss module, and an evaluation module.

feature synthesis for FSOD is shown in Fig. 1, with training divided into three phases.

In the first phase, features from D_B and D_N are extracted using the pre-trained FSOD model F_n to obtain a feature dataset.

In the second training phase, the conditional generator G is trained using real visual features f, their labels y, random noise vector $Z \sim N(0,1)$, and semantic embedding A. Once trained, G can synthesize novel class visual features $f_P \in R^{1024}$ and pseudo-labels y_P based on the given semantic embedding and noise vector. The differentiated synthesized features are obtained by G. The optimization objective for the baseline model's feature generator [12] is:

$$L_G = \min_G \max_D L_{WGAN} + L_C + L_{div} \tag{1}$$

The objective function (1) minimizes L_G with respect to G and maximizes L_WGAN (cw-GAN loss) [13], L_C (to enhance feature synthesis capacity), and L_{div} (to improve feature diversity).

The third phase focuses on training the novel classifier ϕ_{n-cls} separately using synthetic visual features and employing PME loss to efficiently utilize correct and incorrect synthetic features. Finally, the novel classifier ϕ_{n-cls} is concatenated with the base classifier ϕ_{b-cls} from the pre-trained model to obtain the final classifier.

Limited real features for novel classes result in chaotic feature distribution and low-quality features. To address these issues, the synthesized features are constrained and evaluated in 2.2.

2.2 Feature Synthesis

To overcome the limited sample challenge in training for novel classes in FSOD, we propose incorporating semantic embedding information and contrastive loss

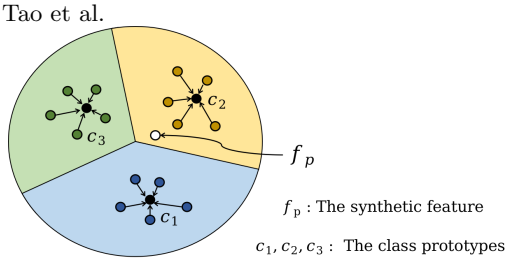

Fig. 2: Illustration of the contrastive loss with synthetic features.

during feature synthesis. However, the feature synthesis process can be unstable, resulting in low-quality synthetic features and pseudo-labels. To address this issue, we introduce a contrastive loss module that utilizes real features as supervision to improve the distribution of synthetic features.

The first step in the feature synthesis process involves calculating class prototypes, which serve as the feature distribution centers for each class. The prototypes reflect the differences between class features and are used to supervise the generator in synthesizing novel class visual features. During training, the synthetic features are guided to be closer to the prototype of their respective class and farther from prototypes of other classes. Similar to other self-supervised studies [14], [15], [16], the contrastive loss in this paper follows formula (2), where n denotes the batch size and C_s denotes the prototypes of the class to which the current generative feature f_{p_i} belongs. Therein, τ is a temperature hyperparameter, which is set to 0.05 by default.

$$L_C = \sum_{i=1}^{n} -\log \frac{\exp\left(f_{p_i} \cdot C_s/\tau\right)}{\sum_{j=1}^{k} \exp\left(f_{p_i} \cdot C_j/\tau\right)} \tag{2}$$

By effectively utilizing different class feature information through contrastive loss, the synthesized visual features are pulled away from other novel class prototypes as well as base class prototypes. The class prototypes provide well-constrained guidance for the synthesized features. The total loss of the generator is defined as formula (3), with α set to 0.1 in the experiments.

$$L_{gan} = L_G + \alpha \cdot L_C \tag{3}$$

L_C, guided by class prototypes, enables the generator to learn realistic feature distribution relationships, resulting in more clustered synthetic features for each class. After training the generator with the total loss L_{gan}, the generator takes the novel class semantic embedding A_N and random noise vector Z as input and generates a set of novel class visual features $\widetilde{f}_p$ and their corresponding pseudo-labels $\widetilde{y}_p$ during inference, as shown in formula (4).

$$\left(\widetilde{f}_p, \widetilde{y}_p\right) = G\left(A_N, Z\right) \tag{4}$$

The synthesized novel class visual features effectively address the challenges in the few-shot object detection task and can be directly used to train the novel class classifier. However, considering the instability of the synthesized features

and pseudo-labels, we comprehensively evaluate them before training the classifier in the subsequent subsection.

2.3 Synthetic Feature Assessment

While the feature distributions are constrained during synthesis, there is no guarantee on the reliability of these features. To tackle this problem, we propose a method to assess the reliability of synthetic features and pseudo-labels using pseudo-margin evaluation loss (PME). By establishing a quality margin between low-quality and high-quality features, we aim to reduce the impact of low-quality features on the classifier.

The uncertainty score s_i^j for each feature is calculated using a pre-trained novel class classification head ϕ_{n-cls} and its cross-entropy loss, see formula (5). Smaller scores indicate more reliable features. These scores are then grouped by category, creating sets of scores S_j for each class j. The PME loss pulls apart the distance between the feature with the largest score and that with the smallest score in each group, ensuring reliable features are far from unreliable ones. This process is repeated for all classes and averaged, as shown in formula (6).

$$s_i^j = \emptyset_{n-cls}\left(f_{p_i}^j, y_{p_i}^j\right) \tag{5}$$

$$L_M = \frac{1}{N}\sum_{j=1}^{N}\max\left(0, \lambda - \max\left(S_j\right) + \min\left(S_j\right)\right) \tag{6}$$

The parameter λ in formula (6) represents the desired spacing between the largest and smallest scores and is set to 0.7 in the experiments. The PME loss aims to create a stable spacing between these scores, aiding the classifier in accurately distinguishing between true and false features. The synthetic feature evaluation loss allows for better utilization of low-quality features compared to training the classifier directly with synthetic features.

The new class classifier is trained using the cross-entropy loss of Faster R-CNN, as shown in equation (7).

$$L_{cls} = -\sum_{j=1}^{N}\sum_{i=1}^{m} y_{p_i}^j \cdot \log s_i^j \tag{7}$$

Finally, the cross-entropy loss is combined with the PME loss to form the total loss of the new class classifier, as shown in formula (8):

$$L_{total} = L_{cls} + L_{cls} \tag{8}$$

The new classifier only includes the novel class classification head ϕ_{n-cls} from the pre-trained FSOD model, and the final classifier is obtained by concatenating ϕ_{n-cls} with the base class classification head ϕ_{b-cls}.

Table 1: The mAP of novel classes on MS-COCO (%).

Sample Num	Method	nAP	nAP50	nAP75	Sample Num	nAP	nAP50	nAP75
10	TFA w/cos [4]	10.0	19.1	9.3	30	13.7	24.9	13.4
	QA-FewDet [20]	10.2	20.4	9.0		16.5	31.9	15.5
	N-PME [21]	10.6	21.1	9.4		14.1	26.5	13.6
	CGDP+FSCN [22]	11.3	23.0	9.8		15.1	29.4	-
	SRR-FSD [6]	11.3	-	9.8		14.7	-	13.5
	FSCE [19]	11.9	-	10.5		16.4	-	16.2
	FSCE*	11.7	**24.3**	9.9		16.4	31.4	16.2
	PDE [23]	12.0	22.3	11.1		**17.2**	31.3	**16.6**
	SVD [24]	12.0	-	10.4		16.0	-	15.3
	FADI [25]	12.2	-	**11.9**		16.1	-	15.5
	BC-YOLO [26]	9.0	-	-		12.9	-	-
	Ours	**12.3**	23.9	11.5		17.1	**31.6**	15.8

3 Experiment

MS-COCO dataset [17]: The proposed method is evaluated on the widely used MS-COCO, which consists of 80 categories, with 60 base classes and 20 novel classes for the few-shot object detection task. Training data for base classes has sufficient annotated instances, while novel classes have only k=10 or k=30 annotated instances per category. The test set includes 5000 images covering both base and novel class instances, with nAP, nAP50, and nAP75 as the common evaluation metrics.

Implementation details: The algorithm is implemented using the MMDetection framework [18], and the pre-trained model of the FSCE [19] algorithm is used as the baseline model. Real visual features of base and novel classes are extracted based on candidate regions with specific IoU thresholds. The semantic vectors are extracted using the Text-Encoder of the CLIP model, which has the same dimension as the noise vectors.

3.1 Comparison with SOTA

Table 1 presents the comparison results of our algorithm with state-of-the-art FSOD algorithms on the MS-COCO dataset. Our method achieves the highest mAP when compared to all other methods on 10 sampled nAP. It also outperforms the second-best method, CGDP+FSCN, in terms of nAP50, with a 1.7 improvement in nAP75. When compared with SRR-FSD [6], which also incorporates semantic information, our method shows significant improvements in both nAP and nAP75. Additionally, compared to the meta-learning-based algorithm QA-FewDet [20], SFC achieves substantial improvements in all three indicators, demonstrating the effectiveness and superiority of our approach. Moreover, our approach surpasses the YOLO-based model BC-YOLO [26] by 3.3% and 4.2% on the 10-shot and 30-shot novel sets, respectively.

Figure 3 visually demonstrates the test results of our method. The detection performance is satisfactory for most categories; however, there is room for improvement in detecting persons due to the presence of unlabeled person data in the base dataset. Overall, our proposed method combining feature synthesis and semantic information proves effective in addressing FSOD tasks and provides new insights into synthetic sample applications.

3.2 Ablation Experiment

We conduct ablation experiments on two modules to validate their effectiveness: Prototype Contrastive Loss and Pseudolabel Distance Evaluation Loss (PME). The baseline model for these experiments is FSCE*. The experiments are performed on the MS-COCO 10-shot dataset, and the results are recorded in Table 2.

From the table, it can be observed that the Prototype Contrastive Loss module significantly enhances the quality of synthesized features generated by the generator. This enhancement is evident in the increased values of nAP and nAP75 by 0.4 and 0.9, respectively. These results demonstrate the effectiveness of guiding the generator to synthesize features based on category prototypes.

Furthermore, the PME module effectively leverages the uncertainty associated with synthetic features, resulting in improvements in nAP, nAP50, and nAP75 by 0.2, 0.7, and 0.7 respectively when compared to models using only contrastive loss. Taking into account the uncertainty of synthetic features proves to be a valid approach.

Table 2: Ablation for key components in MS-COCO 10-shot settings.

CL	SFS	nAP	nAP50	nAP75
-	-	11.7	24.3	9.9
✓	-	12.1	23.2	10.8
✓	✓	12.3	23.9	11.5

Visual analysis of the visual features output by the generator, performed using the t-SNE method [27], further confirms the effectiveness of the Prototype Contrastive Loss module. Figure 4 shows the visual feature distributions for five categories (car, boat, motorcycle, bicycle, and train) before and after applying contrastive loss. The contrastive loss guided by class prototypes improves the distinctiveness of visual features between different classes and enhances the aggregation of visual features within the same class.

In terms of semantic information selection, we compare the performance of CLIP and Fasttext word embeddings. The experiments are conducted on the MS-COCO dataset for both 10-shot and 30-shot settings. Table 4 reveals that CLIP semantic embedding achieves better results in both settings.

Additionally, we explore the influence of different threshold values when mining pseudolabels from the base dataset. The experiments are performed with

Fig. 3: The visualization comparison between ours and the ground truth.

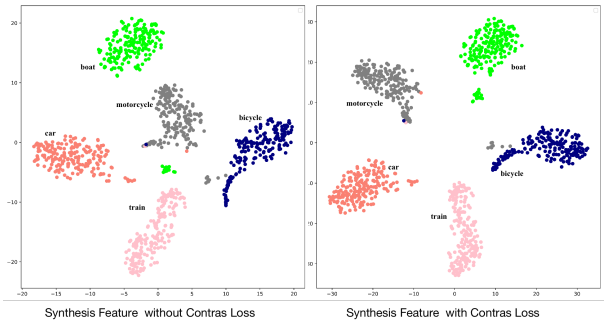

Fig. 4: The Visualization of synthetic features before and after adding contrastive loss.

threshold values of 0.6 and 0.8 on the MS-COCO 30-shot dataset. Table 4 shows that a higher threshold value of 0.8 achieves better accuracy. Lower threshold values introduce more instability and noise to the pseudolabels, resulting in unsatisfactory performance. However, higher threshold values provide more stable and reliable pseudolabel data, leading to performance improvements.

4 Conclusion

This paper introduces a few-shot object detector based on feature synthesis, which can effectively tap the intrinsic connection between semantic information

Table 3: Ablation experiment of semantic information on MS-COCO(%).

shot	semantic	nAP	nAP50	nAP75
10	Fasttext [28]	12.0	24.1	10.8
	CLIP [11]	12.3	23.9	11.5
30	Fasttext [28]	16.8	29.8	15.6
	CLIP [11]	17.1	31.6	15.8

Table 4: Ablation experiments with different threshold pseudo-labels .

	nAP	nAP50	nAP75
Ours	17.1	31.6	15.8
Ours+Pseudo-label(0.6)	17.0	32.4	15.3
Ours+Pseudo-label(0.8)	17.2	33.8	14.6

and visual features to synthesize the visual features of the novel class. In the feature synthesis stage, the proposed algorithm adopts the class prototypes of real visual features as the constraints of the generator, which ensures the quality and distribution of the synthesized visual features. When training the classifier, we fully consider the uncertainty of the synthesized visual features and eliminate the low-quality visual features to improve the performance of the classifier. Ultimately, our method allows for end-to-end training and achieves better results on the MS-COCO detection dataset.

References

1. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9577–9586 (2019)
2. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
3. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(3), 3090–3106 (2022)
4. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)
5. Wu, A., Han, Y., Zhu, L., Yang, Y.: Universal-prototype enhancing for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9567–9576 (2021)
6. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8782–8791 (2021)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)

11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
12. Hayat, N., Hayat, M., Rahman, S., Khan, S., Zamir, S.W., Khan, F.S.: Synthesizing the unseen for zero-shot object detection. In: Proceedings of the Asian Conference on Computer Vision (2020)
13. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10275–10284 (2019)
14. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems 33, 9912–9924 (2020)
15. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020)
16. Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, S., Stiefelhagen, R., Van Gool, L.: Vi2clr: Video and image for visual contrastive learning of representation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1502–1512 (2021)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
18. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
19. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7352–7362 (2021)
20. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3263–3272 (2021)
21. Liu, W., Wang, C., Yu, S., Tao, C., Wang, J., Wu, J.: Novel instance mining with pseudo-margin evaluation for few-shot object detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2250–2254. IEEE (2022)
22. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-shot object detection via classification refinement and distractor retreatment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15395–15403 (2021)
23. Chen, S., Wang, C., Liu, W., Ye, Z., Deng, J.: Pseudo-label diversity exploitation for few-shot object detection. In: International Conference on Multimedia Modeling. pp. 289–300. Springer (2023)
24. Wu, A., Zhao, S., Deng, C., Liu, W.: Generalized and discriminative few-shot object detection via svd-dictionary enhancement. Advances in Neural Information Processing Systems 34, 6353–6364 (2021)
25. Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D.: Few-shot object detection via association and discrimination. Advances in neural information processing systems 34, 16570–16581 (2021)
26. Xia, R., Li, G., Huang, Z., Meng, H., Pang, Y.: Bi-path combination yolo for real-time few-shot object detection. Pattern Recognition Letters 165, 91–97 (2023)

27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(11) (2008)
28. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)