Proceedings of Machine Learning Research vol 278:1–8, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Few-Shot Object Detection via Decoupled and Balanced Contrastive Learning

Pengxin Kang* K_ANG_TX@163.COM
Southwest Institute of Technical Physics, Chengdu 610041, China
Southwest Jiaotong University, Chengdu 611756, China

Yuyong Cui
Southwest Institute of Technical Physics, Chengdu 610041, China

Xiaohe Cao
Southwest Institute of Technical Physics, Chengdu 610041, China

Dong Li
Southwest Institute of Technical Physics, Chengdu 610041, China

Zhenbao Luo
Southwest Institute of Technical Physics, Chengdu 610041, China

Huhai Jiang
Southwest Institute of Technical Physics, Chengdu 610041, China

Qi Zhang
Southwest Institute of Technical Physics, Chengdu 610041, China

Yi Shi
Southwest Institute of Technical Physics, Chengdu 610041, China

Zhonghe Tang
Southwest Institute of Technical Physics, Chengdu 610041, China
Corresponding author

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Abstract. Only a few training examples are used for object detection task, and The productivity of the neural network model will show a dramatic decrease. Countless approaches to few-shot object detection (FSOD) have been formulated to solve the problem with a fine-tuning mechanism. Whereas those methods usually result in misclassification of novel classes and are biased in favor of base classes. For coping with this problem, we advance a fine-tuning learning framework with decoupled and balanced contrastive schemes(FSDB). More precisely, we first incorporate supervised contrastive learning with a decoupled loss to obtain a more outstanding performance for novel classes. Based on the decoupled supervised contrastive learning, we then put forward a class-balanced learning technique to resolve the issue of unequal sample distribution of base and new classes in the fine-tuning procedure. Rigorous experiments conducted on PASCAL VOC and MS-COCO datasets indicate that the presented technique has obtained excellent results for FSOD tasks.

Keywords: Decoupled Contrastive Learning, Class Balanced Learning, Few-shot Detection

1. Introduction

Tremendous advancements have been achieved by convolutional neural networks in the domain of object detection (Lin et al., 2020; Tian et al., 2019). Nonetheless, they rely heavily on abundant training examples. In some actual settings, securing a vast quantity of labeled samples proves to be incredibly arduous,which holds back the outcome of the object detection task. By contrast, humans have the ability to learn patterns from a few samples. With the aim of minimizing the gap separating the performance of deep learning from human capacities, The field of few-shot object detection has been brought to the forefront of research (Qiao et al., 2021; Radford et al., 2021). Currently, the majority of FSOD methods apply the meta-learning methodology to achieve a transfer of task-specific knowledge from base categories to unseen classes. However, these approaches often require a complicated training procedure and data processing, which leads to limited application scenes. In contrast to other methodologies, fine-tuning oriented FSOD approaches exhibit remarkable simplicity and efficiency. Employing a two-tiered fine-tuning training framework, these techniques can attain performance levels comparable to those of meta-learning algorithms, as evidenced in reference (Wang et al., 2020). However, when the majority of model parameters are pre-trained on base datasets and subsequently kept fixed during the adaptation to novel data, potential issues such as suboptimal feature representations and the exacerbation of data imbalance among base classes and unseen classes may arise.

Under the paradigm of fine-tuning based learning, a biphasic training protocol is typically employed. During the opening phase, the object detection network is optimized on the base-class dataset to learn generic features. At the latter stage of the pipeline, the model derived from the first stage is refined using a mix of seen-class and unseen-class instances, by adding a newly initialized classifier to detect unseen classes. This training mechanism results in a base class bias and misclassification. FSCE introduces a contrastive learning branch to enhance feature embedding (Sun et al., 2021), yet it does not take into account the sample imbalance and class balance in contrastive learning.

Motivated by decoupled contrastive learning and balanced contrastive learning, we propose a fine-tuning learning framework with decoupled and balanced contrastive schemes(FSDB) based on DeFRCN (Qiao et al., 2021). We improve DeFRCN for few-shot object detection through two distinct modifications: (1) Decoupled contrastive learning. Due to the unequal distribution of positive versus negative samples, the contrastive learning can not obtain robust feature representations resulting in a misclassification. We adopt the decoupled contrastive loss to achieve the balance between positive and negative samples, thereby enhancing the representation effectiveness of the model and improving the classification effect. (2) Balanced contrastive learning. Because of the separation of the base-class and novel-class training workflows, the model shows a bias towards the base categories, and it hinders effective learning of new categories. We follow a class balancing process into contrastive learning to relieve the conflict between novel and base classes.

This study presents a simple yet effective FSOD approach through the integration of decoupled and balanced contrastive representation learning. During the process of migrating the base-category detector to the new-category detector, we integrated the decoupled contrastive loss into RoI feature aggregation head to estimate the distance between positive and negative proposal encodings. Besides, we designed a class balancing process to alleviate distribution shift among classes.

Overall, our research can be characterized by three principal contributions:

(1) In this research, we introduce an innovative fine-tuning frame for FSOD, which can achieve robust object representations and category balance.

(2) To tackle the imbalance in terms of classes and positive-negative samples,we blend decoupled contrastive loss into RoI head and designed a class balancing process to obtain novel class robust learning.

(3) The devised approach can be seamlessly integrated into prevailing fine-tuning oriented FSOD methodologies to facilitate additional performance enhancement. Rigorous evaluations carried out on the PASCAL VOC and MS-COCO datasets have unequivocally demonstrated that the suggested technique has yielded highly promising outcomes in the domain of FSOD tasks.

2. Related Work

2.1. Few-shot Object Detection

FSOD endeavors to identify objects utilizing a small quantity of labeled instances. This research area can be systematically grouped into two principal paradigms: meta-learning based approaches and fine-tuning based methodologies. Approaches which utilize the meta-learning mechanism endeavor to train a class-unspecific learning-to-learn system to promote the transfer of acquired knowledge from the base-stage detector to the detector dealing with novel categories. FSRW is meta-model oriented on YOLOv2 (Kang et al., 2019), which enhances representation of new class features through a re-weighting module to improve object detection. Meta R-CNN primarily emphasizes the attention mechanism applied to the features of the Region of Interest (RoI) head (Yan et al., 2019). However, these approaches often encounter challenges due to the necessity of designing intricate episodic training frameworks. TFA introduced the balanced dataset in few-shot learning (Wang et al., 2020), which is exceeding meta-learning based approaches in performance. FSCE makes use of contrastive branch to obtain robust feature embedding for improving performance, but it failed to handle positive and negative sample balance and category balance well. DeFRCN achieves the decoupling of the non-class specific RPN and the class-oriented RCNN modules through different gradient scaling factors and cosine attribute similarity metrics. However, most of the above methods rely on complex algorithms and easily encounter overfitting. Our method integrates decoupled contrastive loss into proposal embedding to improve the classification performance.Meanwhile, we incorporate a category balancing process simultaneously to diminish the category deviation caused by the phased training of TFA.

2.2. Contrastive Feature Learning

Lately, contrastive feature learning has made substantial progress in the domain of unsupervised learning. (Chen et al., 2020b,a), which enables the model to l uncover the latent features and capture the relationships within data by making similar samples closer and dissimilar samples farther apart. With contrastive learning, the model can distinguish different images in high-level feature space (Chen et al., 2020b,a,c). Supervised contrastive learning utilizes labeled data to explicitly train the model to distinguish between similar and dissimilar instances.Decoupled contrastive learning reaches a new contrastive objective function to achieve performance improvement with a small-size batch training, but for image classification (Yeh et al., 2022). Balanced contrastive learning mainly

addresses the class imbalance in long-tailed visual recognition (Zhu et al., 2022). The aforementioned methods are mainly designed for image categorization tasks.

By extending decoupled and balanced contrastive feature learning to address the challenges of FSOD, we introduce a fine-tuning framework. Extensive benchmark evaluations unequivocally show that our method surpasses the majority of contemporary FSOD algorithms in detection accuracy.

3. Method

3.1. Problem Specification

Building on the problem settings proposed in previous research (Qiao et al., 2021; Radford et al., 2021), we adopt the conventional problem frameworks for FSOD. The corpus of data used for training purposes is composed of a base set, denoted as D_b, and a novel set, denoted as D_n. The base classes C_b feature an abundance of tagged object instances. In contrast, the novel classes C_n have a scarce number of annotations, and $C_b \cap C_n = \emptyset$. Our approach aligns with fine-tuning based methods in FSOD. These methods typically consist of two key stages: the base class training period and the period of finetuning for novel classes.

3.2. Overview

The entire proposed architecture in this study is illustrated in Figure 1. We herein precisely outline a decoupled and balanced contrastive scheme(FSDB) for FSOD. Our method first applies decoupled contrastive loss to produce proposals contrastive encoding with a small-size batch training, which can better discriminate between novel classes and base classes with more robust features. Next, We employ the category complement in proposals encoding to guarantee that each training batch encompasses all classes. Finally, We adopt the class averaging process in contrastive loss calculation to achieve category balance. All the methods proposed above are applied in the fine-tuning stage.

3.3. Decoupled Contrastive Learning

The supervised contrastive loss has been applied to FSOD, but it ignores the balance problem of positive and negative samples.Inspired by decoupled contrastive learning for classification and recognition in self-supervised learning, we formulate our DCL loss in the manner detailed below with considerations tailored for object detection. Specifically, we denote $\{z_i, y_i\}_{i=1}^{N}$ as the N RoI box features of a mini-batch, where z_i is the feature of the i-th candidate bounding box after being encoded by the contrastive head. The contrastive head is constructed with a deep perceptron architecture. And y_i symbolizes the label associated with the ground truth.

$$L_{dcl} = \frac{-1}{N_{y_i-1}} \left(\sum_{j=1}^{N} \mathbb{I}\{y_i = y_j\} \cdot (\_1) \cdot w(z_i, z_j) \left(\langle z_i, z_j \rangle / \tau \right) + \mathbb{I}\{y_i \neq y_j\} \cdot L_{z_i} \right) \tag{1}$$

$$L_{z_i} = \log \sum_{j=1}^{N} \exp\left(\langle z_i, z_j \rangle \tau \right) \tag{2}$$

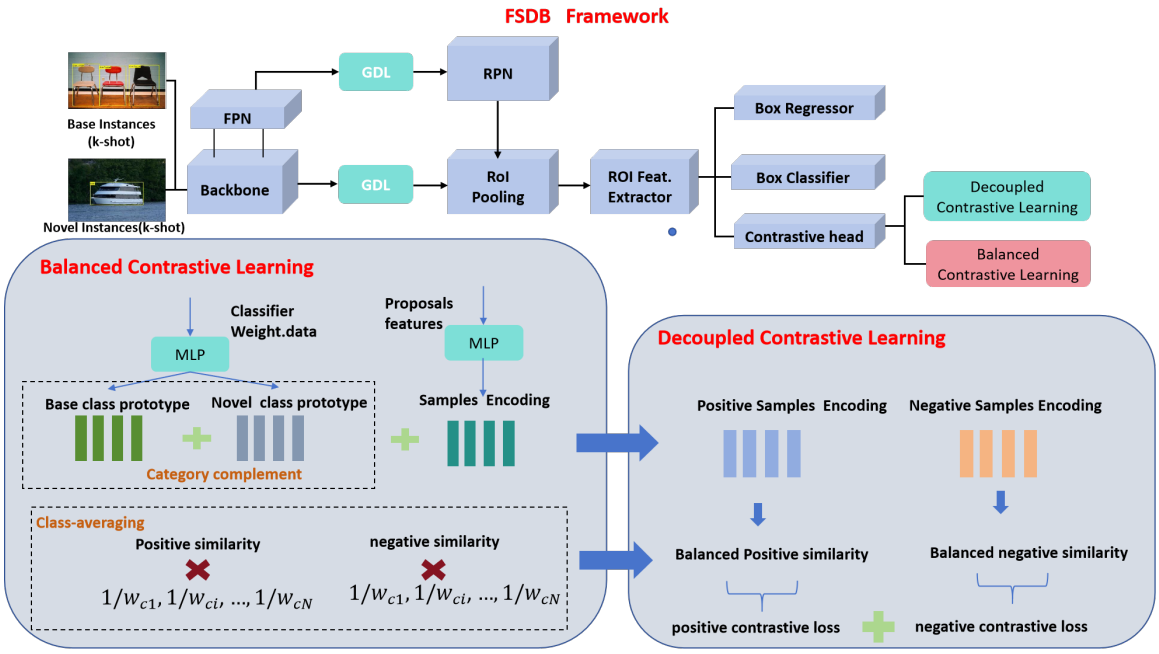

Figure 1: The architecture of FSDB for FSOD. Compared to the DeFRCN, there is a decoupled contrastive learning module and a balanced contrastive learning is inserted into the framework to obtain novel class robust learning. w_{ci} demonstrates the numerical value of instances per classification.

w is a negative von Mises-Fisher weighting function.

$$w\left(z_i, z_j\right) = 2 - \frac{\exp\left(\left\langle z_i, z_j \right\rangle / \tau\right)}{E_i\left[\left(\exp\left(\left\langle z_i, z_j \right\rangle / \tau\right)\right)\right]}, E[w] = 1 \tag{3}$$

3.4. Balanced Contrastive Learning

3.4.1. CATEGORY COMPLEMENT

In order to ensure the presence of all classes within each small batch, we incorporate category-wise prototypes for balanced contrastive learning. Specifically, after applying a nonlinear transformation via a multi - layer perceptron (MLP), we obtain class-specific weights denoted as $w_1, w_2, .., w_k$, which serve as prototypes $z_{c1}, z_{c2}, .., z_{ck}$. To make sure the feature space is a unit hypersphere, we apply L_2 normalization to all representations used in contrastive learning.

3.4.2. CLASS-AVERAGING

The fundamental concept underlying our approach is to calculate the mean of instances within each class for every mini-batch. As a result, this ensures that each class provides approximately an equal contribution to the optimization procedure. From an intuitive perspective, this mechanism serves to mitigate the dominance of base classes in the denominator of the loss function, thereby accentuating

the significance of novel classes. Therefore, we redefine the contrastive loss as follows:

$$L_{dcl} = \frac{-1}{N_{y_i} - 1} \left(\sum_{j=1}^{N} \mathbb{I}\{y_i = y_j\} \cdot \frac{-w(z_i, z_j)(\langle z_i, z_j \rangle / \tau)}{|B_i|} + \mathbb{I}\{y_i \neq y_j\} \cdot BL_{z_i} \right) \tag{4}$$

$$BL_{z_i} = \log \sum_{j=1}^{N} \frac{\exp(\langle z_i, z_j \rangle \tau)}{|B_i|} \tag{5}$$

where the $|B_i|$ term will be subtracted by one when the positive class is averaged.

4. Experiment

4.1. Experimental Configuration

We primarily carry out experiments on Pascal VOC and MS-COCO datasets for FSOD. To maintain the integrity and fairness of the comparative analysis, we strictly adhere to the data partitions and labels established by prior studies (Wang et al., 2020) . Regarding the PASCAL VOC dataset, we segment its 20 classes into three distinct clusters. More precisely, each cluster comprises 15 base classes, along with 5 novel classes. In the case of the COCO dataset, we specify 60 mutually exclusive categories with those in PASCAL VOC as base classes, and the remaining 20 categories act as novel classes.

4.2. Assessment Criteria

When carrying out the experiments on PASCAL VOC dataset experiments, we determined the Average of Precision Values at an intersection over union threshold of 0.5. This enabled us to obtain separate values for base classes (bAP) and novel classes (nAP). By contrast, experiments with the COCO dataset required calculating and reporting the mean Average Precision. This calculation was executed across an IoU range of 0.5 to 0.95, with the analysis restricted to novel classes (nAP).

4.3. Detailed Implementation Protocols

We implement our method using DeFRCN with a ResNet101 backbone trained beforehand on ImageNet. We utilize the Stochastic Gradient Descent (SGD) optimizer and set the batch size to 8 and the learning rate to 1e-4. In the case of PASCAL VOC , the shots settings are $K = \{1,\ 2,\ 3,\ 5,\ 10\}$, we fine-tune for $\{800,\ 1600,\ 2400,\ 3200,\ 4000\}$ iterations, for COCO with $K = \{\ 10,\ 30\}$ shots, we fine-tune for $\{10000,\ 20000\}$ iterations.

4.4. Main Results

4.4.1. PASCAL VOC

As shown in Table 1, FSDB significantly outperforms most of existing FSOD methods. FSDB achieves the best or second-best results on all settings.It is clearly demonstrated by this phenomenon that FSDB remains unbiased with respect to any specified class sets. Moreover, it holds significant potential for generalization to a wider array of common scenarios. FSDB obtains a 58.0% average score and surpasses the second-best result by 1.9%, which demonstrates its effectiveness.

Table 1: Results on the PASCAL VOC

Method / Shots	Novel Set 1					Novel Set 2					Novel Set 3					Avg
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FSRW (Kang et al., 2019)	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9	28.4
Meta R-CNN (Yan et al., 2019)	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1	31.1
TFA (Wang et al., 2020)	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
FSCE (Sun et al., 2021)	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5	46.6
DeFRCN (Qiao et al., 2021)	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4	51.9
VFA (Han and et al., 2023)	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9	56.1
FSDB	**61.3**	**63.4**	**66.1**	**68.4**	**68.5**	**39.0**	**46.1**	**53.1**	**55.1**	**53.9**	**55.9**	**58.5**	**56.9**	**61.4**	**62.8**	**58.0**

4.4.2. COCO

As shown in Table 2, FSDB achieves the best normalized average precision (nAP) among all methods. Our approach outperforms the second-ranked method by approximately 2.5% and 3.8% under 10-shot and 30-shot conditions, respectively, thus validating its robust generalization capability in few-shot object detection. Notably, despite having more trainable parameters than other TFA methods, our approach avoids severe over-fitting, demonstrating the effectiveness of our parameter optimization strategy. In addition, our method has achieved an 11.4% improvement compared with FSCE under the condition of small-batch training, further demonstrating the effectiveness of decoupled contrastive learning.

Table 2: Results on the COCO

Method / Shots	10 shot	30 shot
FSRW (Kang et al., 2019)	5.9	9.1
Meta R-CNN[8]	8.7	12.4
TFA (Wang et al., 2020)	10.0	13.8
FSCE (Sun et al., 2021)	11.9	16.4
DeFRCN (Qiao et al., 2021)	18.5	22.6
VFA (Han and et al., 2023)	16.2	18.9
FSDB(Ours)	**18.7**	**22.7**

5. Conclusion

This paper revisits decoupled and balanced contrastive schemes in fine-tuning learning based FSOD and proposes decoupled contrastive learning and balanced contrastive learning. decoupled contrastive proposal encoding loss can tackle the imbalance in terms of classes and positive-negative samples. Balanced contrastive learning loss can obtain novel class robust learning. Systematic evaluations carried out on the VOC and COCO datasets verify the feasibility of the introduced approach.

References

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020a. doi: 10.48550/arXiv.2006.10029.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020b. doi: 10.48550/arXiv.2002.05709.

X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c. doi: 10.48550/arXiv.2003.04297.

J. Han and et al. Few-shot object detection via variational feature aggregation. In *AAAI*, 2023. doi: 10.1609/aaai.v37i1.25153.

B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00851.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *ICCV*, 2021. doi: 10.48550/arXiv.2108.09017.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. doi: 10.48550/arXiv.2103.00020.

B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, 2021. doi: 10.48550/arXiv.2103.05950.

Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00972.

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. doi: 10.48550/arXiv.2003.06957.

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 *ICCV*, 2019. doi: 10.1109/ICCV.2019.00967.

C.-H. Yeh et al. Decoupled contrastive learning. In *European Conference on Computer Vision*. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19809-0_38.

J. Zhu et al. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00678.