Mutually Reinforcing Structure with Proposal Contrastive Consistency for Few-Shot Object Detection

Tianxue Ma[1], Mingwei Bi[2], Jian Zhang[2], Wang Yuan[1], Zhizhong Zhang[1(✉)], Yuan Xie[1], Shouhong Ding[2], and Lizhuang Ma[1(✉)]

[1] East China Normal University, Shanghai, China
[2] Tencent Youtu Lab, Shanghai, China
{51205901016,51184501076,zzzhang}@stu.ecnu.edu.cn
{mingweibi,timmmyzhang,ericshding}@tencent.com
xieyuan8589@foxmail.com ma-lz@cs.sjtu.edu.cn

Abstract. Few-shot object detection is based on the base set with abundant labeled samples to detect novel categories with scarce samples. The majority of former solutions are mainly based on meta-learning or transfer-learning, neglecting the fact that images from the base set might contain unlabeled novel-class objects, which easily leads to performance degradation and poor plasticity since those novel objects are served as the background. Based on the above phenomena, we propose a Mutually Reinforcing Structure Network (MRSN) to make rational use of unlabeled novel class instances in the base set. In particular, MRSN consists of a mining model which unearths unlabeled novel-class instances and an absorbed model which learns variable knowledge. Then, we design a Proposal Contrastive Consistency (PCC) module in the absorbed model to fully exploit class characteristics and avoid bias from unearthed labels. Furthermore,we propose a simple and effective data synthesis method undirectional-CutMix (UD-CutMix) to improve the robustness of model mining novel class instances, urge the model to pay attention to discriminative parts of objects and eliminate the interference of background information. Extensive experiments illustrate that our proposed approach achieves state-of-the-art results on PASCAL VOC and MS-COCO datasets. Our code will be released at https://github.com/MMatx/MRSN.

Keywords: Few-shot object detection, contrastive learning, data augmentation

1 Introduction

Object detection[48, 9, 12, 51], is a classical task in computer vision, which aims to identify and localize objects in an image. In recent years, the application of deep convolutional neural network [1] has accelerated the development of object detection[26, 18, 16, 17, 19, 35, 34, 50, 49]. However, the remarkable performance

of object detection depends on abundant annotated data. Collecting annotated data is time-consuming and labor-intensive. As opposed to that, a few examples are sufficient for humans to learn a new concept. To bridge this gap, we focus on *few-shot object detection* (FSOD), which aims to adapt the model from the base class to the novel class based on a few labeled novel classes examples.

Based upon the base set, almost all methods for solving FSOD simply utilize the labeled base class objects in the base set as supervision information to train a detector. In the process of using base set, past methods overlook an important phenomenon: material neglect, in which unlabeled novel-class instances are explicitly learned as background. Material neglect has the following drawbacks. 1) An incorrect priori is introduced into the model, which limits the plasticity of the model on novel classes. The model explicitly takes all instances of novel-class that are unlabeled as the background when using the base set to train the detector. Due to incorrect supervision, subsequently, the detector is provided with labeled novel samples for adaptation, it is difficult for a small number of labeled novel class samples to correct the error knowledge. 2) Data augmentation has become a factor in training high-performance few-shot resolver. By providing a variety of samples for the model, the model can obtain a more robust representation space. But the current FSOD solution abandons the solution of data augmentation and even turns a blind eye to the existing unlabeled samples, which is undoubtedly a suboptimal solution to FSOD.

Thus, we advocate a new solution that uses data augmentation for resolving FSOD. Specifically, instead of using meta-learning or transfer-learning to obtain an adaptable model in the base set, we make rational use of unlabeled novel class instances and eliminate the negative effects. Inspired by [28], we use a semi-supervised framework to unearth unlabeled novel class instances and introduce a *Mutually Reinforcing Structure Network* (MRSN), which contains a mining model and an absorbed model.

To fully exploit class characteristics and avoid bias from noise labels, we design the *Proposal Contrastive Consistency* (PCC) module in the absorbed model. We use the mining model to unearth the unlabeled novel-class instances and use the absorbed model to learn the mined novel-class instances. The excavated novel instances may have noise. To prevent the consecutively detrimental effect of noisy pseudo-labels, we utilize PCC to keep the consistency of the mining model and the absorbed model by constructing positive and negative sample pairs between the two models. Meanwhile, we utilize PCC at the proposal level to compare the global and local information of the instance simultaneously, which compensates for the scarcity of training data. Intuitively, for classification head in Faster R-CNN, it ensures that the samples of the same category in the feature space are close to each other, and the regions related to different categories are separable. This coarse goal leads to the loss of detailed features that are important for separating similar samples in the learning process. The PCC method we proposed optimizes the features of "instance recognition", retains the information used to identify the subtle details between the classes and the instances.

Moreover, we propose an effective data synthesis method *undirectional-CutMix* (UD-CutMix) to improve the robustness of model mining novel class instances, urge the model to pay attention to class discrimination features and eliminate the interference of background. Extensive experiments show that the proposed framework achieves significant improvements over state-of-the-art methods on PASCAL VOC and MS-COCO.

2 Related Works

2.1 General Object Detection

At present, object detection methods based on deep learning can be divided into two kinds according to different detection processes. One is two-stage approach [2, 32, 22, 8]. The detector first filters out proposals and then carries out regression and classification on them through the network. [32] uses RPN to generate proposals and ROI head to detection. The other is the one-stage object detection method. The core of the one-stage object detection algorithm is regression, which directly divides categories and regresses the border. [30, 31] are based on a single end-to-end network, which completes the input from the original image to the output of object position and category. [27] combines the regression idea and anchor mechanism, extracts the feature map of different scales, and then regresses the multi-scale regional features of each position in the whole map.

Although the one-stage object detection algorithm has faster detection speed, the two-stage algorithm are more reliable in terms of detection accuracy. In the trade-off between accuracy and speed, we pick the Faster R-CNN with higher accuracy as our benchmark framework.

2.2 Few-Shot Object Detection

Compared with general object detection, the scarcity of labeled novel-class data brings huge difficulty to FSOD. The previous FSOD solving methods can be divided into meta-learning [37, 7] based methods and transfer-learning [40, 36, 29] based methods. The idea of the former is that using the samples of support set to learn the class representation, and then use the learned representation to enhance the response of class related areas in the query sample features. [44] aggregates features by reweighting the query features according to the class encoder. [43] performs channel-wise multiplication of the extracted vector of query features and extracted vector of class features. In the method based on transfer-learning,[45] proposes a context transformer, which learns the ability to integrate context knowledge in the source domain and migrates this ability to the target domain to enhance the recognition ability of the detector, to reduce the object confusion in the few-shot scene. [47] migrates shared changes in the base class to generate more diverse novel-class training examples. With the in-depth investigation of FSOD by researchers, some new methods have emerged. [20] proposes a class margin equilibrium approach to optimize the division of feature

space. [33] puts forward a contrastive loss to alleviate the problem of classification error. [21] solves the task of FSOD from the perspective of classification structure enhancement and sample mining.

However, these methods almost ignore the "material neglect", where the image in the base set contains unlabeled novel-class instances.

2.3 Contrastive Learning

Recently, the success of the self-supervised model in various tasks mainly stems from the use of contrastive learning [10, 3–5]. The goal of contrastive representation learning is to construct an embedding space in which similar pairs of samples remain close to each other while dissimilar ones remain far apart. Contrastive can be achieved by learning the similarity and differences of samples. [4] proposes a simclr framework to maximize the consistency between different views of the same image. [5] uses two neural networks to encode the data to construct positive and negative sample pairs, and encodes the image into query vectors and key vectors respectively. In the training process, try to improve the similarity between each query vector and its corresponding key vector, and reduce the similarity with the key vector of other images. The above methods are mainly used in classification tasks.

As far as we know, there is rare contrastive learning method specially designed for FSOD. Under the proposed MRSN, we propose a contrastive-consistency learning method PCC for FSOD.

3 Method

3.1 Problem Definition

Followed by the widely adopted FSOD setting in [39, 44, 38], our framework aims to perform adaptation on novel categories with the aid of a large and labeled base set D_{base} and a small-scale novel support set D_{novel}. Here, the base set $D_{base} = \{(I_i^{base}, C_i^{base}, B_i^{base})\}$ contains abundant annotated instances, where I_i^{base} denotes the i-th image, $C_i^{base} = \{(c_i^{base})_j\}$ is the entire class set in this image, and $B_i^{base} = \{(x_i, y_i, w_i, h_i)_j\}$ indicates the position of each instance j in this image. While the novel set $D_{novel} = \{(I_i^{novel}, C_i^{novel}, B_i^{novel})\}$ shares the same structure with the base set, the class set C^{base} and C^{novel} of those two collections are non-overlapping. More specifically, the ultimate goal of FSOD task is to classify and locate all instances belonging to the novel-class and maintain the performance on base classes.

3.2 Training Phase

The initialization of our MRSN is very important since the MRSN is an iterative process to unearth unlabeled novel-class instances. Therefore, we first introduce the acquisition of the initialization model of the MRSN, and then introduce the learning process of MRSN in detail.

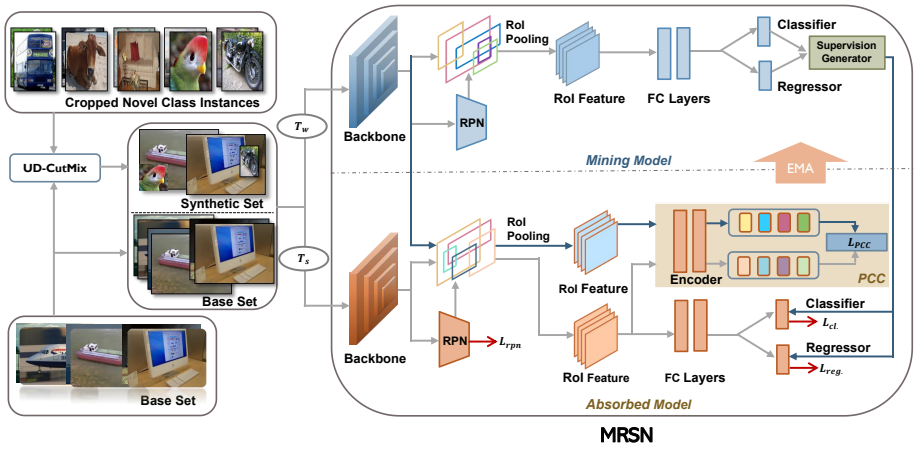

Fig. 1. Framework of our method. First, cropped novel class instances and base set construct a synthetic set through UD-CutMix. Then the images are processed by two different data augmentations (where T_w and T_s denotes different data augmentations) and sent to two models. In MRSN, the outputs generated by the mining model are provided to the absorbed model and the absorbed model transfers the learned knowledge to the mining model through EMA. Meanwhile, to fully exploit class characteristics and avoid bias from noise labels, the PCC module is designed in the absorbed model

Training plain detector. We call the model used to initialize the MRSN as M_{plain}. First, we use D_{base} to train the whole detector. The whole detector $F(\theta_{emb}, \theta_{rpn}, \theta_{roi})$ include the network parameters of feature extraction, RPN module and ROI module, which are expressed as θ_{emb},$\theta_{rpn}$and θ_{roi}, respectively. During the experiment, we found that due to the scarcity of data in novel set, the overfitting problem is easily incurred if we only use the limited data of novel classes to fine-tune the whole detector. To tackle with this problem, we sample a balanced training dataset $D_{balance}$ from D_{base} and D_{novel} as [38]. Specifically for each category in C^{base}, we only take k samples to form $D_{balance}$ together with the whole D_{novel}. The combined training dataset $D_{balance}$ is eventually utilized to fine-tune the classifier and regressor of the detector. In this way, we get a new detector $M_{plain} = F(\theta_{emb}, \theta_{rpn}, \theta^*_{roi})$, which can detect C^{novel} and C^{base} at the same time.

Training mutually reinforcing structure network. To solve material neglect, we propose a method to excavate unlabeled instances of novel classes C^{novel} in base set D_{base}. The specific method is shown in Fig. 1.

Our MRSN is a dual model construction, including M_{mine} and M_{absorb}, where M_{mine} is used to mine unlabeled novel-class instances in D_{base} and give them pseudo labels including corresponding categories and locations, and M_{absorb} is used to learn the mined instances. We utilize the mined instances as explicit supervision for both backbone, RPN and ROIhead. The initialization values

of network parameters of M_{mine} and M_{absorb} are both inherited from $M_{plain} = F(\theta_{emb}, \theta_{rpn}, \theta^*_{roi})$. Based on the above definition, the evolutionary mechanism of MRSN will be introduced in section 3.3. To further impose model discrimination between different categories and avoid the negative impact of noise labels, we design PCC module for M_{absorb} as shown in section 3.4.

3.3 Mutually Reinforcing Structure Network

As illustrated in Fig.1, our MRSN consists of a two-stream detection architecture. We first execute UD-CutMix by recombining labeled novel instances with base images to construct a new synthetic set D_{syn}. Taking images in the D_{syn} and D_{base} with weak data transformation as input, the upper mining stream explores the augmented novel information and further reserves the most credible prediction. The lower absorbed stream gradually adopts the novel knowledge delivered from the upper stream, without forgetting the base knowledge by utilizing base supervision at the same time. Our two-stream framework thus evolves continuously in the process of mutually reinforcing learning.

UD-CutMix. A huge challenge of FSOD is the limited annotations and diversity scarce in the novel set. To address this issue, we propose UD-CutMix combine the cropped novel classes instances and the selected images in D_{base}. UD-CutMix adopts the detector M_{plain} to select the base image, which prediction doesn't contain any novel objects, to be mixed. Specifically, we first crop a novel-class instance from a novel set image, then scale and paste it to a selected base image. By repeating this operation, we can construct a new synthetic set $D_{syn} = \{(I_i^{syn}, C_i^{syn}, B_i^{syn})\}$ as an enlarged and labeled set. The visual comparisons of the original and synthetic images are illustrated in Fig. 2.

The proposed UD-CutMix is distinct from the CutMix [46] in the following two aspects. First, CutMix is agnostic to the category, as it randomly samples data among all categories. UD-CutMix is category-specific, whice samples images in the base dataset and only pastes the novel class instances. We find this strategy can well cope with the lack of novel class data and the imbalance between the base class and novel class labelled data. Second, UD-CutMix crops the complete bounding box of the novel class instance, while CutMix randomly selects a patch to crop, which destroys the instance's global information.

It is notable that some base images involve novel instances. However, those unlabeled novel objects serve as background or even distractors during the base classes training process, hindering the transfer capability to novel classes. We can make full use of these base images by generating pseudo labels. Thus in our framework, the above synthetic set D_{syn} and the original base set D_{base} are both included as supervision in subsequent novel-classes adaptation pipeline.

Novel knowledge mining and absorbing. The novel knowledge mining stream excavates the possible novel instances and assigns them pseudo labels for further supervision.

Given an image $I_i \in D_{base} \cup D_{syn}$ with true label $GT_i = (C_i, B_i)$. We use two different data enhancements for I_i, one is denoted as strong data enhancement T_s, and the other is weak data enhancement T_w. The images applied with T_s are

sent to the absorb stream M_{absorb}, and the images applied with T_w are sent to the mining stream M_{mine}. The corresponding predicted value P_{absorb} and P_{mine} are formulated as the following Equ. 1.

$$\begin{aligned} P_{absorb} =& M_{absorb}(T_s(I_i)) \\ =& \{(c, s, box)_j\}, box = (x, y, w, h), \\ P_{mine} =& M_{mine}(T_w(I_i)) \\ =& \{(c, s, box)_j\}, box = (x, y, w, h), \end{aligned} \tag{1}$$

where c, s and box in $(c, s, box)_j$ respectively represent the category corresponding to the maximum classification probability in the j-th prediction result of image I_i, the corresponding score and the position of the bounding box.

To boost the credibility of pseudo labels and reduce the negative impact caused by noise prediction, we propose a supervision generator to exclude those distractors. We first refer to a confidence score threshold ϕ to get a trusted prediction subset P_{tr}, as expressed in Equ. 2.

$$P_{tr} = \{\mathbb{I}(s \geq \phi)(c, s, box)_j | (c, s, box)_j \in P_{mine}\}, \tag{2}$$

where the filtered prediction P_{tr} is the high-confidence predicted bounding box.

To provide deterministic novel supervision and reduce the learning bias of the base classes, we then set an overlap threshold δ to discard those predictions whose location is at the neighborhood of ground truth, as expressed in Equ. 3.

$$\begin{aligned} S_j &= MAX(IOU(box_j, box_k)), j \in P_{tr}, k \in GT_i, \\ P_{final} &= \{\mathbb{I}(S_j \leq \delta)(c, s, box)_j | (c, s, box)_j \in P_{tr}\}, \end{aligned} \tag{3}$$

where the final prediction P_{final} is the subset of those proposals whose overlapping with the GT bounding boxes of base classes is below a certain threshold.

Finally, we choose the images that novel class instances in the final prediction P_{final} as the supervision images. To stabilize the training process and accelerate the convergence speed, we combine the ground truth labels of selected images with P_{final}. In general, our supervision generator contains bounding box-level confidence filtering as well as overlap filtering, image-level instance filtering, label combining.

Interaction between two stream. By applying supervision generator, the most convincing images subset is obtained for subsequent novel knowledge learning. For M_{absorb}, we use those most convincing images as supervisory information to update the whole model by back-propagation. As expressed in Equ. 4, instead of using gradient back propagation to update M_{mine}, we transfer the parameters of M_{absorb} to iterate M_{mine}. In this way, we can alleviate the accumulation effect on the unreliability of pseudo labels.

$$\theta^t_{mine} = \alpha\theta^{t-1}_{mine} + (1 - \alpha)\theta^t_{absorb}, \tag{4}$$

where α is the hyperparameter to balance the update ratio of M_{mine}. Both θ_{mine} and θ_{absorb} include the network parameters of backbone, RPN module, and ROI

module. When M_{absorb} is updated after a certain number of iterations, we use the absorb model parameters θ_{absorb} to update the freeze mining model parameters θ_{mine}. Here, t and $(t-1)$ represent the network parameters at the current time and before updating, respectively.

3.4 Proposal Contrastive Consistency

The self-supervised method [4, 3, 5, 10] performs well in classification, but not in intensive prediction tasks, such as object detection. The main reason is that the current self-supervised method is global-level feature extraction, and the prediction ability of local is not considered. After in-depth research, we propose a contrastive learning method Proposal Contrastive Consistency (PCC) suitable for our MRSN in FSOD. We introduce a PCC branch to the primary RoI head, parallel to the classification and regression branches. In M_{absorb}, RPN takes feature maps as inputs and generates region proposals, then a mini-batch RoIs [32] is sampled for training. We express RoIs sampled by M_{absorb} as $Roi^{absorb} = \{(x, y, w, h)_j\}$. f_{mine} and f_{absorb} are the feature maps obtained by the feature extractors of M_{mine} and M_{absorb} respectively. According to Roi^{absorb}, RoIPooling is performed on f_{mine} and f_{absorb} to obtain the features of the proposals, and then through the identical encoder E, features in the two feature spaces are mapped to the same area.

$$\begin{aligned} z_j^{mine} &= E(RoIPooling(f_{mine}, Roi_j^{absorb})), \\ z_j^{absorb} &= E(RoIPooling(f_{absorb}, Roi_j^{absorb})). \end{aligned} \tag{5}$$

We take the contrastive learning features from the same proposal and different models as positive sample pairs and others as negative sample pairs. And we measure the cosine similarity between two proposal features in the projected hypersphere.

Our Proposal Contrastive Consistency loss L_{PCC} is formulated as Equ. 6:

$$\begin{aligned} L_{PCC} =& \sum_j -log \frac{Pos_j/\tau}{Neg_j^1/\tau + Neg_j^2/\tau + Pos_j/\tau}, \\ Neg_j^1 =& \sum_{k=1}^{N} \mathbb{I}(k \neq j) exp(sim(z_j^{absorb}, z_k^{absorb})/\tau, \\ Neg_j^2 =& \sum_{k=1}^{N} \mathbb{I}(k \neq j) exp(sim(z_j^{mine}, z_k^{absorb})/\tau, \\ Pos_j =& exp(sim(z_j^{mine}, z_j^{absorb})). \end{aligned} \tag{6}$$

Finally, for our absorb stream M_{absorb}, the total loss is as follows:

$$L = L_{rpn} + \lambda_1(L_{cl.} + L_{reg.}) + \lambda_2 L_{PCC}, \tag{7}$$

where the L_{rpn} contains the cross entropy and regression loss of RPN, the middle two terms $L_{cl.}$, $L_{reg.}$ are the focal loss [24] of classification, smoothed-L1 loss of

Table 1. Performance on the PASCAL VOC dataset. We evaluate the performance on three different sets of novel classes

Method / Shot	Novel Set 1					Novel Set 2					Novel Set 3				
	1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot
MetaDet [39]	17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
RepMet [15]	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
CRDR [21]	40.7	45.1	46.5	57.4	62.4	27.3	31.4	40.8	42.7	46.3	31.2	36.4	43.7	50.1	55.6
FRCN-ft [39]	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
Meta R-CNN [44]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA w/ fc [38]	36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
MPSR [42]	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	**35.6**	-	42.3	48.0	49.7
FSCE [33]	32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0
CME [20]	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
DCNet [13]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
UP [41]	43.8	47.8	50.3	55.4	61.7	31.2	30.5	41.2	42.2	48.3	35.5	**39.7**	43.9	50.6	53.5
Ours	**47.6**	**48.6**	**57.8**	**61.9**	**62.6**	**31.2**	**38.3**	**46.7**	**47.1**	**50.6**	35.5	30.9	**45.6**	**54.4**	**57.4**

regression in roi head respectively, L_{PCC} is the proposal contrastive consistency loss. λ_1 and λ_2 are used to balance the loss functions.

4 Experiments

4.1 Datasets

For a fair comparison with previous work, we use MS-COCO [25] and PASCAL VOC [6] benchmarks to verify the effectiveness of our method.

PASCAL VOC. PASCAL VOC contains 20 categories. According to the previous FSOD experimental settings [14], we divide the 20 categories into 15 base classes and 5 novel classes under three different divisions. We take the trainval sets of the 2007 and 2012 as the training set and PASCAL VOC 2007 test set as the evaluation set. For each category in the novel set, k-shot novel instances are sampled. In FSOD scenarios, we set $k = (1, 2, 3, 5, 10)$.

MS-COCO. MS-COCO contains 80 categories, of which 20 categories are identical to PASCAL VOC. We choose 20 categories in the PASCAL VOC dataset as the novel set, and the rest 60 classes as the base set. For FSOD scenarios in MS-COCO, we set $k = (10, 30)$.

4.2 Implementation Details

We use Faster R-CNN [32] as the basic detector, in which ResNet-101 [11] is served as the backbone and a 4-layer Feature Pyramid Network [23] is utilized for boosting the multi-scale learning process. Models are trained with standard SGD optimizer and batch size of 2 in 1 GPU. We set the learning rate, the momentum, and the weight decay to 0.001, 0.9, and 0.0001, respectively.

Table 2. Few-shot object detection performance on MS-COCO. We report the AP and AP75 on the 20 novel categories

Method / Shot	10-shot		30-shot	
	AP	AP75	AP	AP75
MetaDet 2019 [39]	7.1	6.1	11.3	8.1
Meta R-CNN 2019 [44]	8.7	6.6	12.4	10.8
TFA w/fc 2020 [38]	9.1	8.8	12.1	12.0
MPSR 2020 [42]	9.8	9.7	14.1	14.2
Viewpoint [43]	12.5	9.8	14.7	12.2
FSCE 2021 [33]	11.1	9.8	15.3	14.2
CRDR 2021 [21]	11.3	-	15.1	-
UP 2021 [41]	11.0	10.7	15.6	15.7
Ours	**15.7**	**14.8**	**17.5**	**17.9**

4.3 Comparison Experiments

Results on PASCAL VOC. For all three random novel splits from PASCAL VOC, the evaluation results AP50 are presented in Table 1. The experimental results include $k = (1, 2, 3, 5, 10)$ shot under three different base/novel divisions. Our method is greatly improved over the previous method, which shows the effectiveness of our method. Meanwhile, this indicates that focusing on material neglect plays a key role in solving FSOD. Basically, the best results have been achieved in any shots and any splits. Our method makes the performance of split 2 reach a new level.

Results on MS-COCO. Compared with PASCAL VOC, the MS-COCO contains more categories and more instances in each image, which indicates that there is more serious material neglects in the MS-COCO. Therefore, the improvement of our method on MS-COCO is significant. FSOD results of MS-COCO are shown in Table 2. Compared with baseline methods, TFA [38], our method consistently outperforms its performance. In particular, under the settings of 10 shot and 30 shot, our method is 6.6% and 5.4% higher than TFA on AP, respectively. Meanwhile, in 10-shot setting, our proposed methods gain +3.2% AP and +4.1% AP75 above the current SOTA on the 20 novel classes.

4.4 Ablation

We analyze the effectiveness of proposed modules in our method. Unless otherwise specified, the experiments are carried out on the split1 of PASCAL VOC.

Module effectiveness. We analyze the effects of different modules and show the results in Table 3. From Table 3, we can see that the different designed modules are important to improve the effect. MRSN improves the effect by 3% - 7% under different shot. This shows that MRSN can effectively unearth unlabeled novel-class instances in the base set and generate trusted pseudo labels, to provide richer knowledge for the network to learn. We show unlabeled novel-class instances discovered by MRSN in Fig. 5. With the help of PCC, novel class AP50 can be improved by up to 3.2%. The improvement of effect strongly illustrates

Table 3. Ablation study of different modules for FSOD on PASCAL VOC novel classes (split-1). "MRSN" denotes Mutually Reinforcing Structure Network, "PCC" Proposal Contrastive Consistency, and "avg.△" average performance improvements

MRSN	PCC	UD-CutMix	1-shot	2-shot	3-shot	5-shot	10-shot	avg.△
			39.4	41.5	48.4	52.9	53.7	
✓			43.6	44.5	54.3	59.9	59.7	+5.2
✓	✓		44.9	47.7	57.2	61.1	62.1	+7.4
✓	✓	✓	**47.6**	**48.6**	**57.8**	**61.9**	**62.6**	**+8.5**

the importance of imposing model discrimination between different categories and avoiding the negative impact of noise labels. Although UD-CutMix only improves the effectiveness of the model by 0.5% in 10-shot scenario, it can improve the effect by 2.7% in 1-shot scenario, indicating that in the scenario with extremely scarce data, the MRSN has a weak ability to discover novel instance, and UD-CutMix can provide a supplement to the MRSN, allowing the model to locate and identify novel classes of different scales in different backgrounds. We show images obtained by UD-CutMix in Fig. 2.

Fig. 2. Data obtained by UD-CutMix. The first line is labeled novel instance. The second line is the selected base image. And the third line is synthetic data

Proposal contrastive consistency. The Table 4 shows that the PCC has brought a huge performance improvement. We have tried another form of contrastive learning, which we call Same Class Contrastive Consistency (SCCC). SCCC takes the proposals of the same prediction category as positive samples and the proposals of different categories as negative samples. From the result in Table 4, we can see that the combination of SCCC and our MRSN framework is suboptimal. The reason we analyze is that under the MRSN framework, the

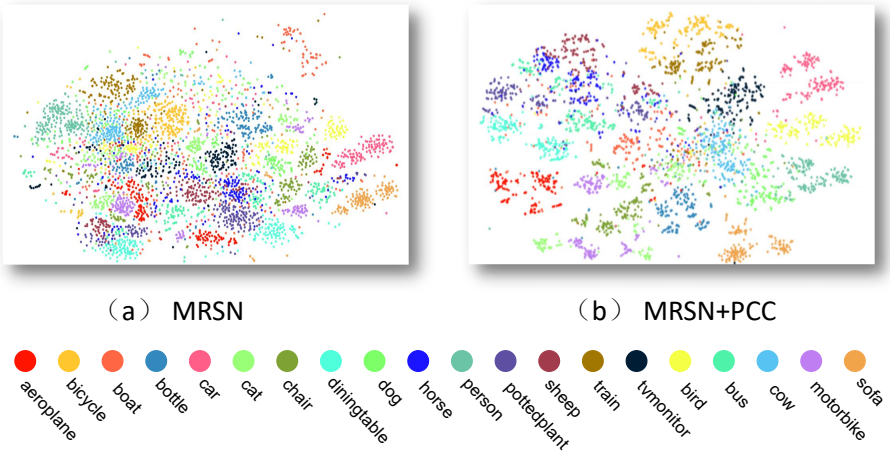

Fig. 3. t-SNE visualization of (a) Object proposal features learned without PCC and (b) Object proposal features learned with PCC

generated pseudo labels have noise. When the category prediction of the proposal deviates, the effect of SCCC is to forcibly close the samples of different categories in the feature space, which will conflict with the optimization goal of the classifier. The PCC regards proposals at the same location as a positive sample pair, and proposals at different locations as a negative sample pair. The learning process is independent of the specific prediction category, which can effectively resist the negative impact caused by noise pseudo labels.

Table 4. Ablation for contrastive learning method, results of novel class AP50

MRSN	UD-CutMix	SCCC	PCC	3-shot	5-shot	10-shot
✓	✓			55.0	60.0	60.2
✓	✓	✓		57.0	60.7	60.9
✓	✓		✓	**57.8**	**61.9**	**62.6**

Fig. 3 shows the object proposal features learned with and without PCC. From t-SNE visualization, we can see that with the help of PCC, the features of different classes are pulled away in the embedded space, while the features of the same class are closer in the embedded space.

Mutually reinforcing structure network. We use M_{mine} in the MRSN to mine unlabeled novel-class instances and generate pseudo label. M_{absorb} learns new knowledge according to the pseudo label provided by M_{mine}, and then M_{mine} accepts the new knowledge learned by M_{absorb} in a certain ratio.

We analyze the influence of the update interval and ratio of M_{mine} on the results. We carry out the experiment under the 3-shot setting of split 1. From

Table 5. Few-shot object detection results on PASCAL VOC base classes (bAP50) under 1,2,3,5,10-shot settings

Method\shot	1-shot	2-shot	3-shot	5-shot	10-shot
MPSR [42]	59.4	67.8	68.4	41.7	51.4
TFA [38]	78.9	78.6	77.3	77.6	75.2
FSCE [33]	72.2	71.8	70.1	74.3	73.6
Ours	**80.2**	**79.7**	**79.3**	**79.5**	**79.2**

the Fig. 4, we can see that the impact of too small update interval and too high update ratio on the M_{mine} is similar, which will reduce the performance. When the interval is too small or the ratio is too high, M_{mine} will shift rapidly to M_{absorb}, which will lead to the drift of the pseudo labels generated by M_{mine}, and reduce the ability of the mutually reinforcing structure to resist noise labels. From the Fig. 4, we can see that when the update interval of M_{mine} is too large, M_{mine} does not receive new knowledge for a long time, which will degrade the performance. In our experiment, we set the update ratio to 0.001 and the update interval to 1000 iterations.

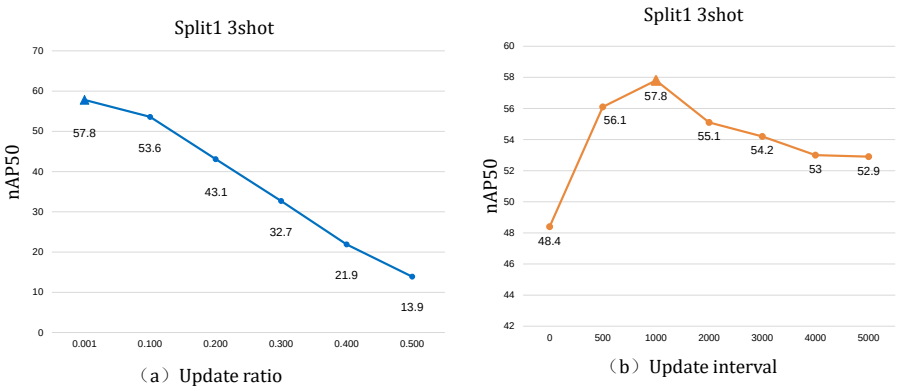

Fig. 4. Ablation study on the mining model update ratio $(1-\alpha)$ and update iterations in the novel classes of the first split of PASCAL VOC

To more intuitively see the role of MRSN, we show the pseudo labels generated by M_{mine} in Fig. 5. From columns (a) and (b), we can see that even in complex scenarios, M_{mine} can mine unlabeled novel-class instances. Columns (c) and (d) show that M_{mine} can give full play to its mining ability in difficult-to-detect scenes (dark) and multi-object scenes. In column (e), the unlabeled novel-class instance is a cow, but M_{mine} is not confused by a similar object horse, and it accurately classifies and locates the cow.

Not forgetting. The previous works almost focus on the performance of the novel classes, while ignoring the non-forgetting effect on the base class. In the real scene, we prefer the model to adapt to the novel classes on the premise that the detection ability of the base class is basically unchanged. The method in this paper achieves the above ideal goal. We show the bAP50 of the model on the base class in Table 5.

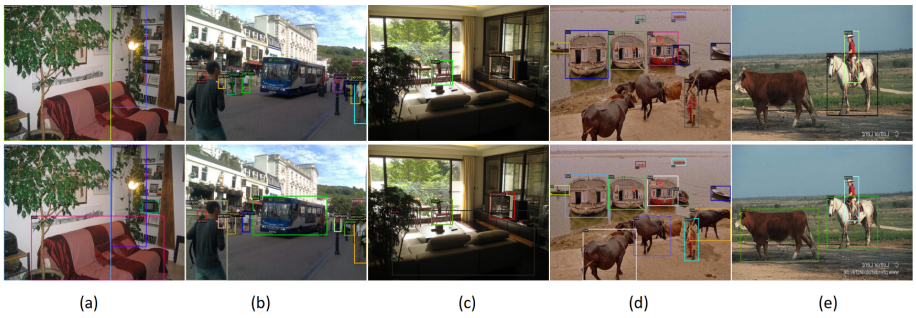

Fig. 5. Visualization of unearthed novel-class instances. The first line is the images from the base set where the novel instances are treated as background, such as "sofa" in columns (a) and (c), "bus" in column (b), "cow" in columns (d) and (e). The second line is the images processed by our mining model

5 Conclusions

In this paper, we discover that previous FSOD solving methods overlook material neglect . Towards solving this problem, a Mutually Reinforcing Structure Network (MRSN) is introduced to mine and absorb unlabeled novel instances. We design a Proposal Contrastive Consistency (PCC) module in MRSN to help learn more detailed features and resist the influence of noise labels. Meanwhile, we design a data synthesis method undirectional-CutMix (UD-CutMix) that combines the novel class instances and the images in the base set. Our method can effectively solve the problem of material neglect in FSOD. Experiments show that our method outperforms the previous methods by a large margin.

Acknowledgment: This work is supported by National Key Research and Development Program of China (2019YFC1521104, 2021ZD0111000), National Natural Science Foundation of China (72192821, 61972157, 62176092, 62106075), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Art major project of National Social Science Fund (I8ZD22), Shanghai Science and Technology Commission (21511101200, 22YF1420300, 21511100700), Natural Science Foundation of Shanghai (20ZR1417700), CAAI-Huawei MindSpore Open Fund.

References

1. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 international conference on engineering and technology (ICET). pp. 1–6. Ieee (2017)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
8. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
9. Gu, Q., Zhou, Q., Xu, M., Feng, Z., Cheng, G., Lu, X., Shi, J., Ma, L.: Pit: Position-invariant transform for cross-fov domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8761–8770 (2021)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
11. 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)
12. He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., Liu, W., Tong, Y., Ma, L., Zhang, L.: End-to-end video object detection with spatial-temporal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1507–1516 (2021)
13. Hu, H., Bai, S., Li, A., Cui, J., Wang, L.: Dense relation distillation with context-aware aggregation for few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10185–10194 (2021)
14. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8420–8429 (2019)
15. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2019)
16. Li, B., Sun, Z., Guo, Y.: Supervae: Superpixelwise variational autoencoder for salient object detection. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.

pp. 8569–8576. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33018569, https://doi.org/10.1609/aaai.v33i01.33018569
17. Li, B., Sun, Z., Li, Q., Wu, Y., Hu, A.: Group-wise deep object co-segmentation with co-attention recurrent neural network. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 8518–8527. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00861, https://doi.org/10.1109/ICCV.2019.00861
18. Li, B., Sun, Z., Tang, L., Sun, Y., Shi, J.: Detecting robust co-saliency with recurrent co-attention neural network. In: Kraus, S. (ed.) Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. pp. 818–825. ijcai.org (2019). https://doi.org/10.24963/ijcai.2019/115, https://doi.org/10.24963/ijcai.2019/115
19. Li, B., Sun, Z., Wang, Q., Li, Q.: Co-saliency detection based on hierarchical consistency. In: Amsaleg, L., Huet, B., Larson, M.A., Gravier, G., Hung, H., Ngo, C., Ooi, W.T. (eds.) Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019. pp. 1392–1400. ACM (2019). https://doi.org/10.1145/3343031.3351016, https://doi.org/10.1145/3343031.3351016
20. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond max-margin: Class margin equilibrium for few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7363–7372 (2021)
21. 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)
22. 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)
23. 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)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
25. 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: European conference on computer vision. pp. 740–755. Springer (2014)
26. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. International journal of computer vision **128**(2), 261–318 (2020)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
28. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: International Conference on Learning Representations (2021)
29. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)

30. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
31. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
33. 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)
34. Tang, L., Li, B.: CLASS: cross-level attention and supervision for salient objects detection. In: Ishikawa, H., Liu, C., Pajdla, T., Shi, J. (eds.) Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part III. Lecture Notes in Computer Science, vol. 12624, pp. 420–436. Springer (2020). https://doi.org/10.1007/978-3-030-69535-4_26, https://doi.org/10.1007/978-3-030-69535-4_26
35. Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 3560–3570. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00356, https://doi.org/10.1109/ICCV48922.2021.00356
36. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242–264. IGI global (2010)
37. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial intelligence review **18**(2), 77–95 (2002)
38. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)
39. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-learning to detect rare objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9925–9934 (2019)
40. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 1–40 (2016)
41. Wu, A., Han, Y., Zhu, L., Yang, Y., Deng, C.: Universal-prototype augmentation for few-shot object detection. arXiv preprint arXiv:2103.01077 (2021)
42. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: European Conference on Computer Vision. pp. 456–472. Springer (2020)
43. Xiao, Y., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. In: European Conference on Computer Vision. pp. 192–210. Springer (2020)
44. 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)
45. Yang, Z., Wang, Y., Chen, X., Liu, J., Qiao, Y.: Context-transformer: tackling object confusion for few-shot detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12653–12660 (2020)
46. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF

International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6022–6031. IEEE (2019)
47. Zhang, W., Wang, Y.X.: Hallucination improves few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13008–13017 (2021)
48. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
49. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4504–4513 (June 2022)
50. Zhong, Y., Li, B., Tang, L., Tang, H., Ding, S.: Highly efficient natural image matting. CoRR **abs/2110.12748** (2021), `https://arxiv.org/abs/2110.12748`
51. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: End-to-end video object detection with spatial-temporal transformers. arXiv preprint arXiv:2201.05047 (2022)