Dense Relation Distillation with Context-aware Aggregation for Few-Shot Object Detection

Hanzhe Hu[1], Shuai Bai[2], Aoxue Li[1], Jinshi Cui[1]*, Liwei Wang[1]*
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]Beijing University of Posts and Telecommunications
{huhz, lax}@pku.edu.cn baishuai@bupt.edu.cn {cjs, wanglw}@cis.pku.edu.cn

Abstract

Conventional deep learning based methods for object detection require a large amount of bounding box annotations for training, which is expensive to obtain such high quality annotated data. Few-shot object detection, which learns to adapt to novel classes with only a few annotated examples, is very challenging since the fine-grained feature of novel object can be easily overlooked with only a few data available. In this work, aiming to fully exploit features of annotated novel object and capture fine-grained features of query object, we propose Dense Relation Distillation with Context-aware Aggregation (DCNet) to tackle the few-shot detection problem. Built on the meta-learning based framework, Dense Relation Distillation module targets at fully exploiting support features, where support features and query feature are densely matched, covering all spatial locations in a feed-forward fashion. The abundant usage of the guidance information endows model the capability to handle common challenges such as appearance changes and occlusions. Moreover, to better capture scale-aware features, Context-aware Aggregation module adaptively harnesses features from different scales for a more comprehensive feature representation. Extensive experiments illustrate that our proposed approach achieves state-of-the-art results on PASCAL VOC and MS COCO datasets. Code will be made available at https://github.com/hzhupku/DCNet.

1. Introduction

With the success of deep convolutional neural works, object detection has made great progress these years [20, 23, 8]. The success of deep CNNs, however, heavily relies on large-scale datasets such as ImageNet [2] that enable the training of deep models. When the labeled data becomes scarce, CNNs can severely overfit and fail to generalize. While in contrast, human beings have exhibited strong performance in learning a new concept with only a few examples available. Since some object categories naturally have scarce examples or bounding box annotations are laborsome to obtain such as medical data. These problems have triggered increasing attentions to deal with learning models with limited examples. Few-shot learning aims to train models to generalize well with a few examples provided. However, most existing few-shot learning works focus on image classification [29, 26, 27] problem and only a few focus on few-shot object detection problem. Since object detection not only requires class prediction, but also demands localization of the object, making it much more difficult than few-shot classification task.

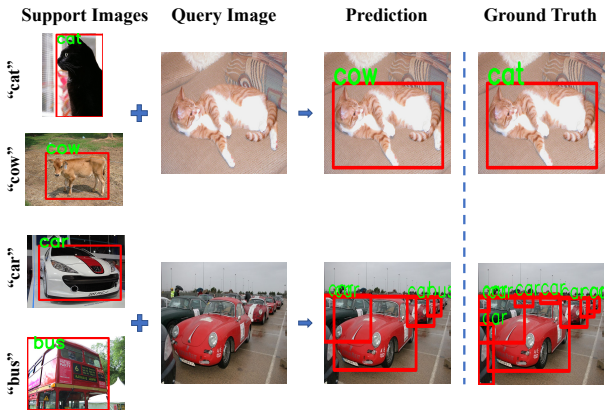

Figure 1. Two challenges for few-shot object detection. a) Appearance changes between support and query images are common, which results in a misleading manner. b) Occlusion problem brings about incomplete feature representation, causing false classification and missing detection.

Prior studies in few-shot object detection mainly consist of two groups. Most of them [13, 35, 34] adopt a meta-learning [5] based framework to perform feature reweighting for a class-specific prediction. While Wang *et al.* [31] adopt a two-stage fine-tuning approach with only fine-tuning the last layer of detectors and achieve state-of-the-art

* Corresponding authors.

performance. Wu *et al.* [33] also use similar strategy and focus on the scale variation problem in few-shot detection.

However, aforementioned methods often suffer from several drawbacks due to the challenging nature of few-shot object detection. Firstly, relations between support features and query feature are hardly fully explored in previous few-shot detection works, where global pooling operation on support features is mostly adopted to modulate the query branch, which is prone to loss of detailed local context. Specifically, appearance changes and occlusions are common for objects, as shown Fig. 1. Without enough discriminative information provided, the model is obstructed from learning critical features for class and bounding box predictions. Secondly, although scale variation problem has been widely studied in prior works [17, 15, 33], it remains a serious obstacle in few-shot detection tasks. Under few-shot settings, feature extractor with scale-aware modifications is inclined to overfitting, leading to a deteriorated performance for both base and novel classes.

In order to alleviate the above issues, we first propose the dense relation distillation module to fully exploit support set. Given a query image and a few support images from novel classes, the shared feature learner extracts query feature and support features for subsequent matching procedure. Intuitively, the criteria that determines whether query object and support object belong to the same category mainly measures how much feature similarity they share in common. When appearance changes or occlusions occur, local detailed features are dominant for matching candidate objects and template ones. Hence, instead of obtaining global representations of support set, we propose a dense relation distillation mechanism where query and support features are matched in a pixel-wise level. Specifically, key and value maps are produced from features, which serve as encoding visual semantics for matching and containing detailed appearance information for decoding respectively. With local information of support set effectively retrieved for guidance, the performance can be significantly boosted, especially in extremely low-shot scenarios.

Furthermore, for the purpose of mitigating the scale variation problem, we design the context-aware feature aggregation module to capture essential cues for different scales during RoI pooling. Since directly modifying feature extractor could result in overfitting, we choose to perform adjustment from a more flexible perspective. Recognition of objects with different scales requires different levels of contextual information, while the fixed pooling resolution may bring about loss of substantial context information. Hence, an adaptive aggregation mechanism that allocates specific attention to local and global features simultaneously could help preserve contextual information for different scales of objects. Therefore, instead of performing RoI pooling with one fixed resolution, we choose three different pooling resolutions to capture richer context features. Then an attention mechanism is introduced to adaptively aggregate output features to present a more comprehensive representation.

The contributions of this paper can be summarized as follows:

1. We propose a dense relation distillation module for few-shot detection problem, which targets at fully exploiting support information to assist the detection process for objects from novel classes.

2. We propose an adaptive context-aware feature aggregation module to better capture global and local features to alleviate scale variation problem, boosting the performance of few-shot detection.

3. Extensive experiments illustrate that our approach has achieved a consistent improvement on PASCAL VOC and MS COCO datasets. Specially, our approach achieves better performance than the state-of-the-art methods on the two datasets.

2. Related Work

2.1. General Object Detection

Deep learning based object detection can be mainly divided into two categories: one-stage and two-stage detectors. One-stage detector YOLO series [20, 21, 22] provide a proposal-free framework, which uses a single convolutional network to directly perform class and bounding box predictions. SSD [18] uses default boxes to adjust to various object shapes. On the other hand, RCNN and its variants [7, 9, 6, 23, 8] fall into the second category. These methods first extract class-agnostic region proposals of the potential objects from a given image. The generated boxes are then further refined and classified into different categories by subsequent modules. Moreover, many works are proposed to handle scale variance [17, 15, 24, 25]. Compared to one-stage methods, two-stage methods are slower but exhibit better performance. In our work, we adopt Faster R-CNN as the base detector.

2.2. Few-Shot Learning

Few-shot learning aims to learn transferable knowledge that can be generalized to new classes with scarce examples. Bayesian inference is utilized in [4] to generalize knowledge from a pretrained model to perform one-shot learning. Meta-learning based methods have been prevalent in few-shot learning these days. Metric learning based methods [16, 29, 26, 27] have achieved state-of-the-art performance in few-shot classification tasks. Matching Network [29] encodes input into deep neural features and performs weighted nearest neighbor matching to classify query images. Our proposed method is also based on matching mechanism. Prototypical Network [26] represents each

class with one prototype which is a feature vector. Relation Network [27] learns a distance metric to compare the target image with a few labeled images. While optimization based methods [19, 5] are proposed for fast adaptation to new few-shot task. [11] proposes a cross-attention mechanism to learn correlations between support and query images. Above methods are focusing on the few-shot classification task while few-shot object detection problem is relatively under-explored.

2.3. Few-Shot Object Detection

Few-shot object detection aims to detect object from novel classes with only a few annotated training examples provided. LSTD [1] and RepMet [14] adopt a general transfer learning framework which reduces overfitting by adapting pre-trained detectors to few-shot scenarios. Recently, Meta YOLO [13] designs a novel few-shot detection model with YOLO v2 [21] that learns generalizable meta features and automatically reweights the features for novel classes by producing class-specific activating coefficients from support examples. Meta R-CNN [35] and FsDetView [34] perform similar process with base detector as Faster R-CNN. TFA [31] simply performs two-stage finetuning approach by only finetuning the classifier on the second stage and achieves better performance. MPSR [33] proposes multi-scale positive sample refinement to handle scale variance problem. CoAE [12] proposes non-local RPN and focuses on one-shot detection from the view of tracking by comparing itself with other tracking methods, while our method performs cross-attention on features extracted by the backbone in a more straightforward way and targets at few-shot detection task. FSOD [3] proposes attention-RPN, multi-relation detector and contrastive training strategy to detect novel object. In our work, we adopt the similar meta-learning based framework as Meta R-CNN and further improve the performance. Moreover, with our proposed method, the class-specific prediction procedure can be successfully removed, simplifying the overall process.

3. Method

3.1. Preliminaries.

Problem Definition. Following setting in [13, 35], object classes are divided into base classes C_{base} with abundant annotated data and novel classes C_{novel} with only a few annotated samples, where C_{base} and C_{novel} have no intersection. We aim to obtain a few-shot detection model with the ability to detect objects from both base and novel classes in testing by leveraging generalizable knowledge from base classes. The number of instances per category for novel classes is set as k (*i.e.*, k-shot).

We align the training scheme with the episodic paradigm [29] for few-shot scenario. Given a k-shot learning task, each episode is constructed by sampling: 1) a support set containing image-mask pairs for different classes $S = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{h\times w\times 3}$ is an RGB image, $y_i \in \mathbb{R}^{h\times w}$ is a binary mask for objects of class i in the support image generated from bounding box annotations and N is the number of classes in the training set; 2) a query image q and annotations m for the training classes in the query image. The input to the model is the support pairs and query image, the output is detection prediction for query image.

Basic Object Detection. The choice of base detectors is varied. [13] utlizes YOLO v2 [21] which is a one-stage detector, while [35] adopts Faster R-CNN [23] which is a two-stage detector and provides consistently better results. Therefore, we also adopt Faster R-CNN as our base detector which consists of a feature extractor, region proposal network (RPN) and the detection head (RoI head).

Feature Reweighting for Detection. We choose Meta-RCNN [35] as our baseline method. Formally, let I denote an input query image, $\{I_{si}, M_{si}\}|_{i=1}^{N}$ denote support images and masks converted from bounding-box annotations, where N is the number of training classes. RoI features $z^j|_{j=1}^{n}$ is generated by the RoI pooling layer (n is the number of RoIs) and class-specific vectors $w_i \in \mathbb{R}^C, i = 1, 2, ..., N$ are produced with a reweighting module which shares its backbone parameters with the feature extractor, where C is the feature dimension. Then class-specific feature z_i is achieved with:

$$z_i = z \otimes w_i, i = 1, 2, ..., N, \tag{1}$$

where $\otimes$ denotes channel-wise multiplication. Then class-specific prediction is performed to output the detection results. Based on this methodology, we further make a significant improvement and simplify the prediction procedure by removing the class-specific prediction.

3.2. DCNet

As illustrated in Fig. 2, we present the Dense Relation Distillation (DRD) module with Context-aware Feature Aggregation (CFA) module to fully exploit support features and capture essential context information. The two proposed components form the final model DCNet. We will first depict the architecture of the proposed DRD module. Then we will bring out the details of the CFA module.

3.2.1 Dense Relation Distillation Module

Key and Value Embedding. Given a query image and support set, query and support features are produced by feeding them into the shared feature extractor. The input of the dense relation distillation (DRD) module is the query feature and support features. Both parts are first encoded into pairs of key and value maps through the dedicated deep en-

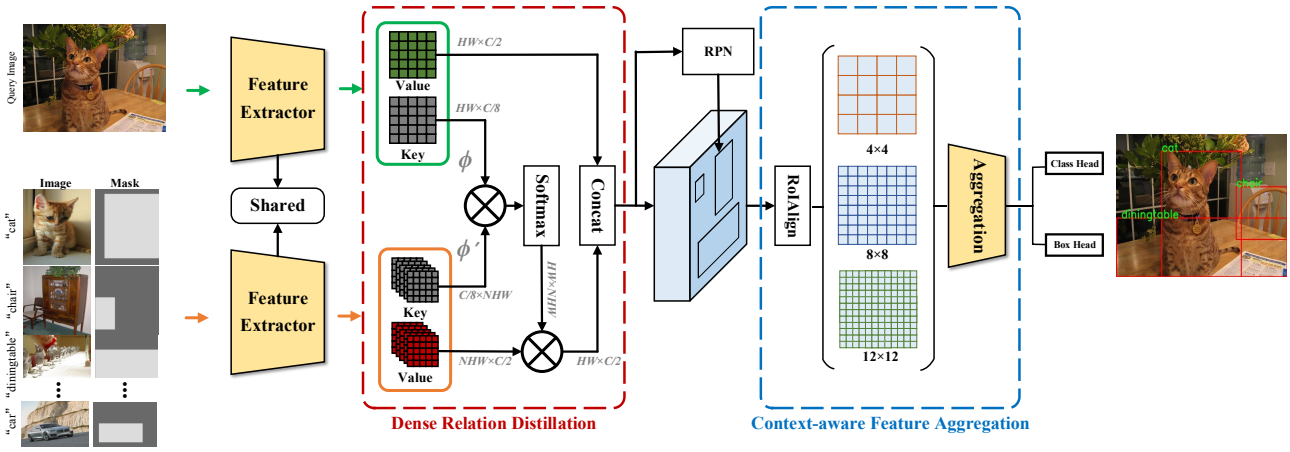

Figure 2. The overall framework of our proposed DCNet. For training, the input for each episode consists of a query image and N support image-mask pairs from N classes. The shared feature extractor first produces query feature and support features. Then, the dense relation distillation (DRD) module performs dense feature match to activate co-exisiting features of input query. With proposals produced by RPN, context-aware feature aggregation (CFA) module adaptively harnesses features generated with different scales of pooling operations, capturing different levels of features for a more comprehensive representation.

coders. The query encoder and support encoder adopt the same structure while not sharing parameters.

The encoder takes one or multiple feature as input and outputs two feature maps for each input feature: key and value with two parallel 3×3 convolution layers, which serve as reducing the dimension of the input feature to save computation cost. Specifically, key maps are used for measuring the similarities between query feature and support features, which help determine where to retrieve relevant support values. Therefore, key maps are learned to encode visual semantics for matching and value maps store detailed information for recognition. Hence, for query feature, the output is a pair of key and value maps: $k_q \in \mathbb{R}^{C/8\times H\times W}, v_q \in \mathbb{R}^{C/2\times H\times W}$, where C is the feature dimension, H is the height, and W is the width of input feature map. For support features, each of the features is independently encoded into key and value maps, the output is $k_s \in \mathbb{R}^{N\times C/8\times H\times W}, v_s \in \mathbb{R}^{N\times C/2\times H\times W}$, where N is the number of target classes (also the number of support samples). The generated key and value maps are further fed into the relation distillation part where keys maps of query and support are densely matched for addressing target objects.

Relation Distillation. After acquiring the key/value maps of query and support features, relation distillation is performed. As illustrated in Fig. 2, soft weights for value maps of support features are computed via measuring the similarities between key maps of query feature and support features. The pixel-wise similarity is performed in a non-local manner, formulated as:

$$F(\boldsymbol{k_{qi}}, \boldsymbol{k_{sj}}) = \phi(\boldsymbol{k_{qi}})^T \phi'(\boldsymbol{k_{sj}}), \tag{2}$$

where i and j are the index of the query and support location, ϕ, ϕ' denote two different linear transformations with parameters learned via back propagation during training process, forming a dynamically learned similarity function. After computing the similarity of pixel features, we perform softmax normalization to output the final weight W :

$$W_{ij} = \frac{exp(F(\boldsymbol{k_{qi}}, \boldsymbol{k_{sj}}))}{\sum_j exp(F(\boldsymbol{k_{qi}}, \boldsymbol{k_{sj}}))}. \tag{3}$$

Then the value of the support features are retrieved by a weighted summation with the soft weights produced and then it is concatenated with the value map of query feature. Hence, the final output is formulated as:

$$y = concat[v_q, W * v_s], \tag{4}$$

where $*$ denotes matrix inner-product. Noted that there are N support features, which brings N key-value pairs. We perform summation over N output results to obtain the final result, which is a refined query feature, activated by support features where there are co-existing classes of objects in query and support images.

Previous trials [13, 35, 34] utilize class-wise vectors generated by global pooling of support features to modulate the query feature, which guide the feature learning from a holistic view. However, since appearance changes or occlusions are common in natural images, the holistic feature may be misleading when objects of the same class vary much between query and support samples. Also, when most parts of the objects are unseen due to the occlusions, the retrieval of local detailed features becomes substantial, which former methods completely neglect. Hence, equipped with the dense relation distillation module, pixel-level relevant information can be distilled from support features. As long as there exist some common characteristics, the pixels of

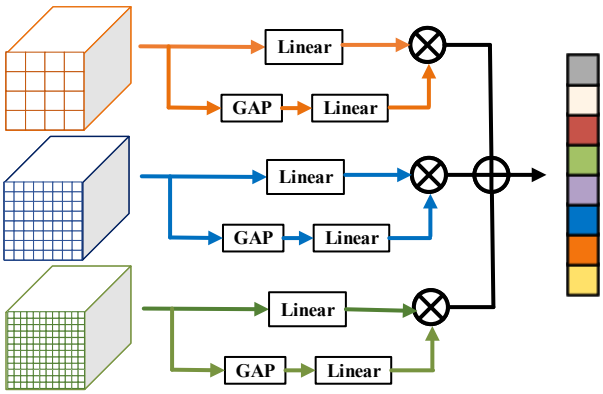

Figure 3. Illustration of context-aware feature aggregation. Attention mechanism is adopted to adaptively aggregate different features, where the weights are normalized with softmax function.

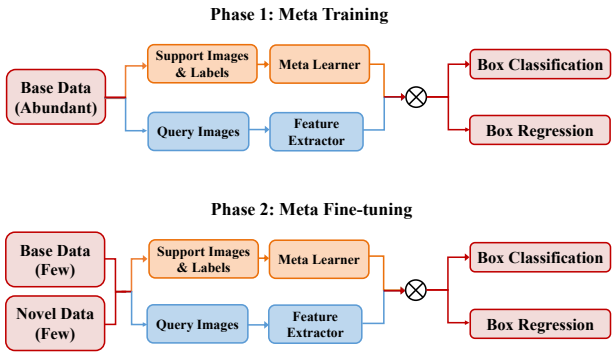

Figure 4. Demonstration of learning strategy of meta-learning based few-shot detection framework. The meta learner aims to acquire meta information and help the model to generalize to novel classes.

query features belonging to the co-existing objects between query and support samples will be further activated, providing a robust modulated feature to facilitate the prediction of class and bounding-box.

Our distillation method can be seen as an extension of the non-local self-attention mechanism [28, 30]. However, instead of performing self-attention, we specially design the relation distillation model to realize information retrieval from support features to modulate the query feature, which can be treated as a cross attention.

3.2.2 Context-aware Feature Aggregation

After performing dense relation distillation, DRD module has fulfilled its duty. The refined query feature is subsequently fed into RPN where region proposals are output. Taking proposals and feature as input, RoI Align module performs feature extraction for final class prediction and bounding-box regression. Normally, pooling operation is implemented with a fixed resolution 8 in our original implementation, which is likely to cause information loss during training. For general object detection, this kind of information loss can be remedied with large scale of training data, while the problem becomes severe in few-shot detection scenarios with only a few training data available, which is inclined to induce a misleading detection results. Moreover, with scale variation amplified due to the few-shot nature, the model tends to lose the generalization ability to novel classes with adequate adaption to different scales. To this end, we propose Context-aware Feature Aggregation (CFA) module. Instead of using a fixed resolution 8, we empirically choose 4, 8 and 12 three resolutions and perform parallel pooling operation to obtain a more comprehensive feature representation. The larger resolution tends to focus on local detailed context information specially for smaller objects, while the smaller resolution targets at capturing holistic information to benefit the recognition of larger objects, providing a simple and flexible way to alleviate the scale variation problem.

Since each generated feature contains different level of semantic information. With the intention to efficiently aggregate features generated from different scales of RoI pooling, we further propose an attention mechanism to adaptively fuse the pooling results. As illustrated in Fig. 3, we add an attention branch for each feature which consists of two blocks. The first block contains a global average pooling. The second one contains two consecutive fc layers. Afterwards, we add a softmax normalization to the generated weights for balancing the contribution of each feature. Then the final output of the aggregated feature is the weighted summation of the three features.

3.3. Learning Strategy

As illustrated in Fig. 4, we follow the training paradigm in [13, 35, 34], which consists of meta-training and meta fine-tuning. In the phase of meta-training, abundant annotated data from base classes is provided. We jointly train the feature extractor, dense relation distillation module, context-aware feature aggregation module and other basic components of detection model. In meta fine-tuning phase, we train the model on both base and novel classes. As only k labeled bounding-boxes are available for the novel classes, to balance between samples from base and novel classes, we also include k boxes for each base class. The training procedure is the same as the meta-training phase but with fewer iterations for model to converge.

4. Experiments

In this section, we first introduce the implementation details and experimental configurations in Sec. 4.1. Then we present our detailed experimental analysis on PASCAL VOC dataset in Sec. 4.2 together with ablation studies and

qualitative results. Finally, results on COCO dataset will be presented in Sec. 4.3.

4.1. Datasets and Settings

Following the instructions in [13], we construct the few-shot detection datasets for fair comparison with other state-of-the-art methods. Moreover, to achieve a more stable few-shot detection results, we perform 10 random runs with different randomly sampled shots. Hence, all the results in the experiments is averaged results by 10 random runs.

PASCAL VOC. For PASCAL VOC dataset, we train our model on the VOC 2007 trainval and VOC 2012 trainval sets and test the model on VOC 2007 test set. The evaluation metric is the mean Average Precision (mAP). Both the trainval sets are split by object categories, where 5 are randomly chosen as novel classes and the left 15 are base classes. We use the same split as [13], where novel classes for four splits are {"bird", "bus", "cow", "motorbike" ("mbike"), "sofa"}, {"aeroplane"("aero", "bottle", "cow", "horse", "sofa"}, {"boat", "cat", "motorbike", "sheep", "sofa"}, respectively. For few-shot object detection experiments, the few-shot dataset consists of images where k object instances are available for each category and k is set as 1/3/5/10.

COCO. MS COCO dataset has 80 object categories, where the 20 categories overlapped with PASCAL VOC are set to be novel classes. 5000 images from the validation set noted as minival are used for evaluation while the left images in the train and validation set are used for training. The process of constructing few-shot dataset is similar to PASCAL VOC dataset and k is set as 10/30.

Implementation Details. We perform training and testing process on images with a single scale. The shorter side of the query image is resized to 800 pixels and longer sides are less than 1333 pixels while maintaining the aspect ratio. The support image is resized to a squared image of 256×256. We adopt ResNet-101 [10] as feature extractor and RoI Align [8] as RoI feature extractor. The weights of the backbone is pre-trained on ImageNet [2]. After training on base classes, only the last fully-connected layer (for classification) is removed and replaced by a new one randomly initialized. It is worth noting that all parts of the model participate in learning process in the second meta fine-tuning phase without any freeze operation. We train our model with a mini-batch size as 4 with 2 GPUs. We utilize the SGD optimizer with the momentum of 0.9, and weight decay of 0.0001. For meta-training on PASCAL VOC, models are trained for 240k, 8k, and 4k iterations with learning rates of 0.005, 0.0005 and 0.00005 respectively. For meta fine-tuning on PASCAL VOC, models are trained for 1300, 400 and 300 iterations with learning rates as 0.005, 0.0005 and 0.00005 respectively. As for MS COCO dataset, during meta-training, models are trained for 56k, 14k and 10k iterations with learning rates of 0.005, 0.0005 and 0.00005 respectively. And during meta fine-tuning, model are trained for 2800, 700 and 500 iteration for 10-shot fine-tuning and 5600, 1400 and 1000 iterations for 30-shot fine-tuning.

Baseline Method. Since we adopt Faster-RCNN as base detector, we choose Meta R-CNN [35] as the baseline method. Moreover, we implement it by ourselves for a more fair comparison.

4.2. Experiments on PASCAL VOC

In this section, we conduct experiments on PASCAL VOC dataset. We first compare our method with the state-of-the-art methods. Then we carry out ablation studies to perform comprehensive analysis of the components of our proposed DCNet. Finally, some qualitative results are presented to provide an intuitive view of the validity of our method. For all the experiments, we run 10 trials with random support data and report the averaged performance.

4.2.1 Comparisons with State-of-the-art Methods

In Table 1, we compare our method with former state-of-the-art methods which mostly report results with multiple random runs. Our proposed DCNet achieves state-of-the-art results on almost all the splits with different shots and outperforms previous methods by a large margin. Specifically, in extremely low-shot settings (*i.e.* 1-shot), our method outperforms others by about 10% in split 1 and 3, providing a convincing proof that our DCNet is able to capture local detailed information to overcome the variations brought by the randomly sampled training shots.

4.2.2 Ablation Study

We present results of comprehensive ablation studies to analyze the effectiveness of various components of the proposed DCNet. All ablation studies are conducted on the PASCAL VOC 2007 test set with the first novel splits. All results are averaged over 10 random runs.

Impact of dense relation distillation module. We conduct experiments to validate the superiority of the proposed dense relation distillation (DRD) module. Specifically, we implement the baseline method for meta-learning based few-shot detection Meta R-CNN with class-specific prediction for the final box classification and regression. While the DRD module requires no extra class-specific processing. As shown in line 1 and 2 of Table 2, DCNet w/o CFA equals to Faster R-CNN equipped with DRD module, our proposed DRD module achieves consistent improvement on all novel splits with all shots number, which effectively demonstrates the supremacy of the relation distillation mechanism over the baseline method. Moreover, the improvement over baseline is significant when the shot number is low, which proves that the DRD module successfully exploits useful information from limited support data.

	Novel Set 1					Novel Set 2					Novel Set 3				
Methods / Shots	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [1]	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
Meta YOLO [13]	14.8	15.5	26.7	33.9	47.2	15.7	15.2	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet* [32]	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN* [35]	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 [31]	22.9	34.5	40.4	46.7	52.0	16.9	26.4	30.5	34.6	39.7	15.7	27.2	34.7	40.8	44.6
TFA* w/cos [31]	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
FsDetView* [34]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	**31.9**	**37.0**	45.7	21.2	30.0	37.2	**43.8**	49.6
DCNet*(ours)	**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**

Table 1. Few-shot object detection performance on VOC 2007 test set of PASCAL VOC dataset. We report the mAP with IoU threshold 0.5 (AP50) under three different splits for five novel classes. * denotes the results averaged over multiple random runs.

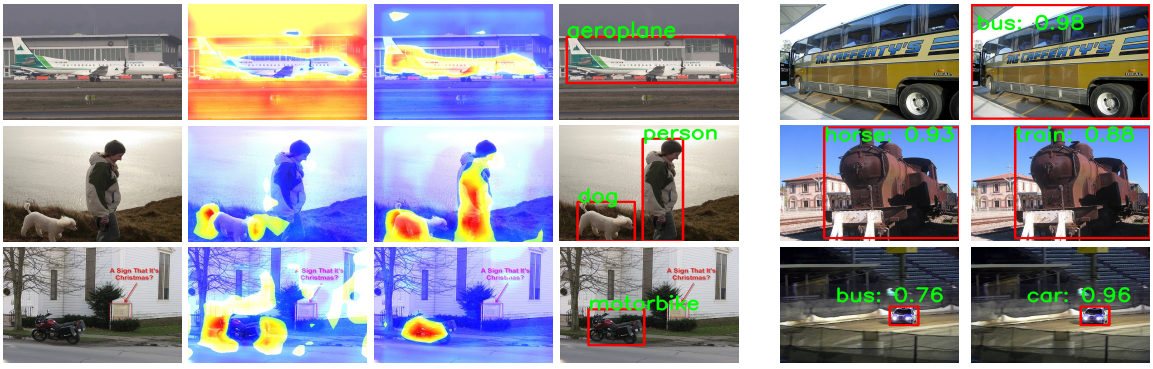

Figure 5. (a). Visualizations of features before and after dense relation distillation module. (b). Visualizations of effect of context-aware feature aggregation module.

	Novel Set 1				
Methods / Shots	1	2	3	5	10
Meta R-CNN†	21.8	27.9	38.6	45.0	51.4
DCNet w/o CFA	31.1	35.9	42.6	48.2	57.2
Meta R-CNN† w/CFA	22.0	31.3	39.6	45.6	52.6
Meta R-CNN† w/CFA*	24.2	32.0	40.2	47.8	53.0
DCNet w/CFA	32.9	36.8	43.1	49.1	57.6
DCNet	**33.9**	**37.4**	**43.7**	**51.1**	**59.6**

Table 2. Ablation study to evaluate the effectiveness of different components in our proposed method. The mAP with IoU threshold 0.5 (AP50) is reported. * denotes CFA module with attention aggregation fashion. † denotes our implementation.

Methods / Resolution	4	8	12	10-shot
DCNet	✓	-	-	56.8
DCNet	-	✓	-	57.2
DCNet	-	-	✓	58.7
DCNet	✓	✓	-	57.9
DCNet	-	✓	✓	59.1
DCNet	✓	-	✓	58.9
DCNet	✓	✓	✓	**59.6**

Table 3. The impact of different RoI pooling resolutions. The experiments are conducted on VOC 2007 test set of PASCAL VOC dataset with novel split1 and AP50 on 10-shot task averaged from 10 random runs is reported.

Impact of context-aware feature aggregation module. We carry out experiments to evaluate the validity of the proposed context-aware feature aggregation (CFA) module. Specifically, RoI features generated from parallel branches are aggregated with a simple summation. From line 1 and 3 of the table, with the introduction of CFA module, Meta R-CNN achieves notable gains over the baseline. Since CFA module targets at preserving detailed information in a scale-aware manner, different levels of detailed features can be retrieved to assist the prediction process.

Impact of different RoI pooling resolutions. To further evaluate the impact of different RoI pooling resolutions, we perform explicit experiments to show the detailed perfor-

Shots	Methods	Average Precision						Average Recall					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
10	LSTD [1]	3.2	8.1	2.1	0.9	2.0	6.5	7.8	10.4	10.4	1.1	5.6	19.6
	Meta YOLO [13]	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
	MetaDet* [32]	7.1	14.6	6.1	1.0	4.1	12.2	11.9	15.1	15.5	1.7	9.7	30.1
	Meta R-CNN* [35]	8.7	19.1	6.6	2.3	7.7	14.0	12.6	17.8	17.9	7.8	15.6	27.2
	TFA* w/fc [31]	9.1	17.3	8.5	-	-	-	-	-	-	-	-	-
	TFA* w/cos [31]	9.1	17.1	8.8	-	-	-	-	-	-	-	-	-
	FsDetView* [34]	12.5	27.3	9.8	2.5	13.8	19.9	20.0	25.5	25.7	7.5	27.6	38.9
	DCNet*(ours)	**12.8**	23.4	**11.2**	**4.3**	**13.8**	**21.0**	18.1	**26.7**	25.6	**7.9**	24.5	36.7
30	LSTD [1]	6.7	15.8	5.1	0.4	2.9	12.3	10.9	14.3	14.3	0.9	7.1	27.0
	Meta YOLO [13]	9.1	19.0	7.6	0.8	4.9	16.8	13.2	17.7	17.8	1.5	10.4	33.5
	MetaDet* [32]	11.3	21.7	8.1	1.1	6.2	17.3	14.5	18.9	19.2	1.8	11.1	34.4
	Meta R-CNN* [35]	12.4	25.3	10.8	2.8	11.6	19.0	15.0	21.4	21.7	8.6	20.0	32.1
	TFA* w/fc [31]	12.0	22.2	11.8	-	-	-	-	-	-	-	-	-
	TFA* w/cos [31]	12.1	22.0	12.0	-	-	-	-	-	-	-	-	-
	FsDetView* [34]	14.7	30.6	12.2	3.2	15.2	23.8	22.0	28.2	28.4	8.3	30.3	42.1
	DCNet*(ours)	**18.6**	**32.6**	**17.5**	**6.9**	**16.5**	**27.4**	**22.8**	27.6	**28.6**	**8.4**	25.6	**43.4**

Table 4. Few-shot object detection performance on COCO minival of MS COCO dataset. We report the mean Averaged Precision and mean Averaged Recall on the 20 novel classes of COCO. * denotes the results averaged over multiple random runs.

mance. As shown in Table 3, solely adopting larger pooling resolution could yield better performance. However, only when aggregating features generated with all three resolutions, the best performance could be obtained.

Impact of attentive aggregation fashion for CFA module. Based on the plain CFA module, we further propose an attention-based aggregation mechanism to adaptively fuse different RoI features. As presented in line 3 and line 4 of Table 2, the attention aggregation mechanism can further boost the performance of the model, which promotes the plain CFA module with a more comprehensive feature representation, effectively balancing the contributions of each extracted features. Finally, with the combination of DRD module and CFA module, we present DCNet, which achieves the best performance according to Table 2.

4.2.3 Qualitative Results

To further comprehend the effect of dense relation distillation (DRD) module, we visualize features before and after DRD module. As shown in Fig. 5 (a), after relation distillation, query features can be activated to facilitate the subsequent detection procedure. Moreover, different from former meta-learning based methods which performs prediction in a class-wise manner, our proposed DRD module can model relations between query and support features in all classes at the same time as shown in the second line of Fig. 5 (a). The DRD module enables the model to focus more on the query objects under the guidance of support information. Additionally, we also visualize the effect of CFA module presented in Fig. 5 (b). With a relatively large or small query object as input, DCNet w/o CFA suffers from false classification or missing detection , while the introduction of CFA module could effectively resolve this issue.

4.3. Experiments on MS COCO

We evaluate 10/30-shot setups on MS COCO benchmark and report the averaged performance with the standard COCO metrics over 10 runs with random shots. The results on novel classes can be seen in Table 4. Despite the challenging nature of COCO dataset with large number of categories, our proposed DCNet achieves state-of-the-art performance on most of the metrics.

5. Conclusions

In this paper, we have presented the Dense Relation Distillation Network with Context-aware Aggregation (DCNet) to tackle few-shot object detection problem. Dense relation distillation module adopts dense matching strategy between query and support features to fully exploit support information. Furthermore, context-aware feature aggregation module adaptively harnesses features from different scales to produce a more comprehensive feature representation. The ablation experiments demonstrate the effectiveness of each component of DCNet. Our proposed DCNet achieves state-of-the-art results on two benchmark datasets, *i.e.* PASCAL VOC and MS COCO.

Acknowledgment

This work was supported by the National Key R&D Program of China under grant 2017YFB1002804.

References

[1] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2836–2843. AAAI Press, 2018.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

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

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4005–4016, 2019.

[12] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *Advances in Neural Information Processing Systems*, pages 2725–2734, 2019.

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

[14] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2019.

[15] Yonghyun Kim, Bong-Nam Kang, and Daijin Kim. San: Learning relationship between convolutional features for multi-scale object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 316–331, 2018.

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

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

[19] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

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

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[24] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018.

[25] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *Advances in neural information processing systems*, pages 9310–9320, 2018.

[26] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[27] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

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

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[31] Xin Wang, Thomas E. Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9919–9928. PMLR, 2020.

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

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

[34] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 192–210. Springer, 2020.

[35] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9577–9586, 2019.