Conditional Prototype Learning for Few-Shot Object Detection

Zhenwei He
hzw@cqut.edu.cn
Xinye Liao
xinyeliao@stu.cqut.edu.cn
Xin Feng*
xfeng@cqut.edu.cn

School of Computer Science and Engineering
Chongqing University of Technology
Chongqing, China

Abstract

Few-Shot Object Detection (FSOD) aims to train models for detecting novel objects by leveraging abundant data from base classes and only a few examples from novel classes. In the meta-learning paradigm for FSOD, the class prototypes are learned to guide the feature learning. However, The particularity of limited samples may lead prototypes to fit specific instances — semantic bias, which compromises the generalization ability. To address this, we propose Conditional Prototype Learning (CPL), a novel framework that integrates robust conditional references with training data to mitigate bias and enhance prototype robustness. Specifically, we introduce a Masked Conditional Variational Autoencoder (MCVA) to generate more robust conditional prototypes by correcting semantic biases. Furthermore, since ideal prototypes should not retain sample-specific particularities, the conditional prototypes lack explicit spatial and scale information. We design the task-aware feature aggregation (TFA) module to enhance the feature presentation for classification and regression tasks. Extensive experiments on PASCAL VOC and MS COCO demonstrate that our approach achieves state-of-the-art performance, validating the effectiveness of CPL in improving generalization to novel classes. Our code is now available at https://github.com/yefeng23323/CPL

1 Introduction

Object detection, as a core task in computer vision, has achieved remarkable progress in recent years driven by advances in deep learning. Despite these achievements, conventional object detection models typically rely on large-scale, meticulously annotated, and uniformly distributed training samples to achieve strong generalization performance. It is a challenge for machines to perform efficient detection using only a few samples. In contrast, human vision can easily recognize new objects from a limited number of examples. To bridge the gap between human and machine vision, few-shot object detection has been proposed as an important subject.

Few-Shot Object Detection (FSOD) addresses the challenge of detecting objects under limited data conditions. It operates under the assumption that base classes have abundant

* Corresponding author.

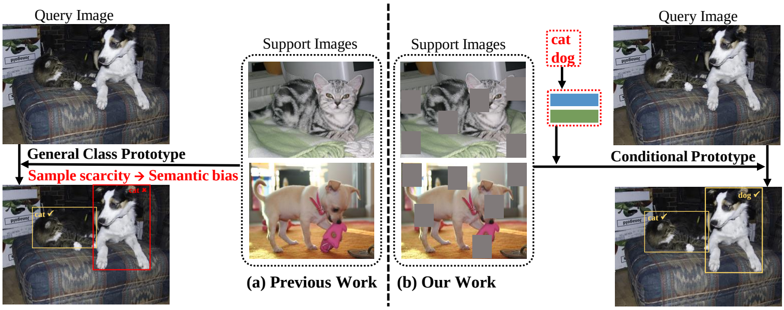

Figure 1: (a) Previous work extracted class prototypes directly from the support images, which are subject to large semantic biases due to the scarcity of samples, leading to misrecognition. (b) We propose conditional prototype learning for FSOD, which implements robust conditional references to mitigate bias and enhance prototype robustness.

training samples, while novel classes are represented by only k examples. Meta-learning has been shown to be an effective paradigm for accomplishing this task. Recently, many meta-learning methods have focused on learning more effective class prototypes. For instance, VFA [14] replaces class prototypes with variational features for aggregation with query features. However, because limited samples may meet similar sample-specific particularities, these methods, which only learn the prototypes from the support samples, may still encounter sample-specific particularities [43], leading to semantic bias in the learned prototypes. To address this limitation, we propose leveraging conditional references to guide prototype learning and mitigate bias. In this paper, we introduce Conditional Prototype Learning (CPL), which incorporates word embeddings as conditional references to enhance training. Specifically, we propose the Masked Conditional Variational Autoencoder (MCVA) to generate more robust conditional prototypes by correcting semantic biases. Additionally, to address the inherent limitations of conditional prototypes, we design a task-aware feature aggregation (TFA) module to enhance feature presentation for classification and regression tasks, ensuring more robust and accurate detection. The Masked Conditional Variational Autoencoder (MCVA) is introduced to refine prototypes using conditional references, addressing two considerations: 1) Since conditional references, such as word embeddings, are free from sample-specific particularities, class prototypes should learn from these references during training to reduce semantic bias. 2) Given the limited samples available for novel class training, class prototypes should be enriched by fusing broader category semantics from the conditional references. Inspired by the success of Masked Autoencoders (MAE) [17], we incorporate a similar masking mechanism to achieve these goals. Specifically, MCVA refines semantic bias and integrates category semantics into class prototypes by masking and restoring them with conditional references. This process ensures that the resulting conditional prototypes are both robust and semantically enriched.

Conventional meta-learning models often rely on class prototypes to enhance features for detection tasks. However, these approaches overlook a critical aspect: detection requires not only category information from prototypes but also scale information derived from sample-specific particularities. To address this, we propose the Task-Aware Feature Aggregation (TFA) module, which implement feature enhancement for the two subtasks of detection—classification and regression. Specifically, TFA enhances scale information by interacting query features with objects of varying sizes, improving the model's ability to localize objects accurately. Simultaneously, it strengthens category information by aggregating conditional prototypes, ensuring classification performance. This approach optimizes the feature representation for classification and regression tasks, respectively. In a word, with both the MCVA and TFA, the proposed CPL can achieve better performance in FSOD task. Our contribution can be summarized as follows:

- We introduce the Masked Conditional Variational Autoencoder (MCVA) to refine the semantic bias in class prototypes, generating more robust conditional prototypes. The model avoids overfitting to sample-specific particularities and achieves greater robustness to novel classes with MCVA.

- Considering that the classification and regression tasks need different kinds of features, we propose the Task-Aware Feature Aggregation (TFA) module, which separately enhances features for the two tasks.

- Extensive experimental results demonstrate that our method significantly improves few-shot object detection performance, achieving state-of-the-art results on the Pascal VOC and COCO datasets.

2 Related Works

2.1 General Object Detection

In recent years, deep learning-based object detection has achieved remarkable progress. Existing object detectors can be broadly categorized into two types: single-stage detectors and two-stage detectors. Single-stage detectors [8, 23, 29, 30, 33] directly predict from CNN feature maps, while two-stage detectors [1, 5, 9, 10, 16, 22, 31] typically generate object proposals first and then perform category classification and box localization. Additionally, anchor-free detectors, such as the DETR series [3, 32, 46, 49], have garnered increasing attention. However, most of these detectors rely on large amounts of training data, which makes them unsuitable for scenarios with limited training samples.

2.2 Few-Shot Object Detection

Few-Shot Object Detection (FSOD) aims to solve the challenge of achieving effective object detection in data-scarce scenarios and has become an important research direction in computer vision. Existing methods are primarily categorized into three types: data augmentation, transfer learning, and meta-learning approaches. **Data augmentation based** [36, 40, 41, 48] methods alleviate the impact of data scarcity by expanding training samples or exploring the intrinsic diversity of data features. **Transfer learning based** [2, 27, 34, 35, 38, 44] methods (also called fine-tuning method) focus on knowledge transfer, employing a two-stage fine-tuning strategy to adapt knowledge learned from large-scale datasets to novel classes with limited samples. **Meta-learning based** [19, 24, 42, 45] methods allow the model to learn how to learn by simulating a sample less learning scenario. Usually two parts, support set and query set, are defined in the task, which are processed in parallel by the support branch and query branch, respectively.

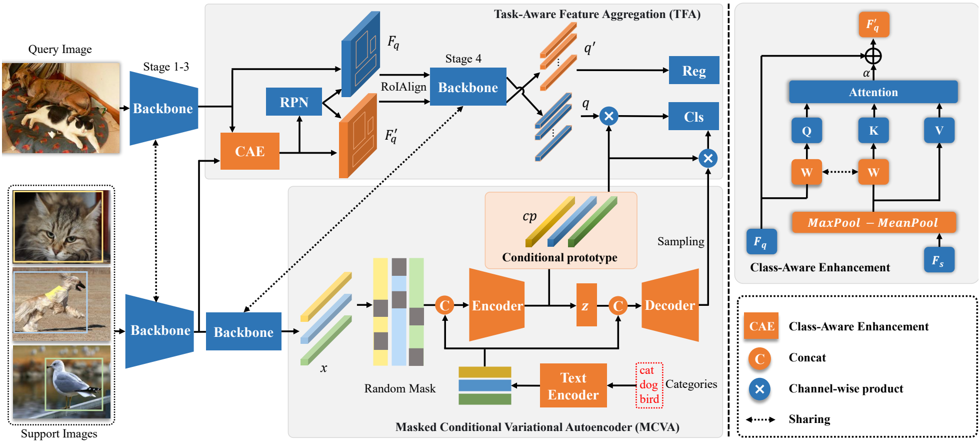

Figure 2: The overall architecture of our Conditional Prototype Learning (CPL). MCVA and TFA are proposed to improve the performance.

3 Methods

In this section, we introduce the proposed Conditional Prototype Learning (CPL) framework for Few-Shot Object Detection (FSOD). First, we outline the problem setup for FSOD and provide an overview of the model architecture. Subsequently, we describe the details of two key components: the Masked Conditional Variational Autoencoder (MCVA) and the Task-Aware Feature Aggregation (TFA).

3.1 Problem Setting

This work adopts the standard problem setting of Few-Shot Object Detection (FSOD) following [35, 42]. Specifically, let $D = \{(x,y)|x \in X, y \in Y\}$, where x represents an input image and y denotes its corresponding instance annotations. The class set C is divided into a base class set C_b and a novel class set C_n, satisfying $C_b \cup C_n = C$ and $C_b \cap C_n = \emptyset$. Each class in C_b contains abundant training data, while each class in C_n includes only $K (K = 1,2,3,...)$ annotated instances. The goal of FSOD is to detect objects from the novel classes C_n by leveraging the abundant instances in C_b along with the K shots instances in C_n.

3.2 The Model Architecture

Our Conditional Prototype Learning (CPL) approach is built on the widely-used Faster R-CNN framework, with ResNet-101 [15] serving as the backbone, as illustrated in Fig. 2. Both support and query images are input into the two-stem backbone, respectively, for meta-learning. The support features are then passed into the Masked Conditional Variational Autoencoder (MCVA), which refines semantic biases and generates conditional prototypes for detection. Simultaneously, the support features are fed into the Task-aware Feature Aggregation (TFA) module to enhance features for regression. After that, the MCVA provides unbiased contional prototypes and reconstructed features into the TFA for classification. By integrating MCVA and TFA, our proposed model achieves improved performance in Few-Shot Object Detection (FSOD).

3.3 Masked Conditional Variational Autoencoder

Under the meta-learning paradigm, class prototypes enable models to quickly adapt to unseen categories. However, in the case of sample-specific particularities and insufficient sample numbers, class prototypes can become highly sensitive to training data, leading to semantic bias. To address this issue, we introduce the Masked Conditional Variational Autoencoder (MCVA) to reduce this bias, and thus generate more robust conditional prototypes. MCVA is based on conditional reference learning and will not be limited by specific samples. In addition, the inclusion of a masking mechanism allows a better fusion of semantic information while also reducing the risk of sample overfitting.

Specifically, MCVA employs a masking and restoration mechanism with conditional references, such as word embeddings. When support images are processed through the backbone network, they yield support features x_i. These features are first randomly masked and then concatenated with their corresponding category word embeddings w_i to produce $\hat{x}_i$:

$$\hat{x}_i = C(w_i, M(x_i)) \tag{1}$$

where $M(\cdot)$ represents the random masking operation, which sets a portion of the feature elements to zero. And $C(\cdot)$ represents concat operation. Following this, we fed the feature $\hat{x}_i$ into the VAE encoder $\mathcal{F}_{enc}$, which produces a latent feature z for feature restoration and a condional prototype cp for detection:

$$\mu_i, \sigma_i = \mathcal{F}_{enc}(\hat{x}_i) \tag{2}$$

$$cp_i = \mu_i + \sigma_i, z_i = \mu_i + \sigma_i \times \varepsilon_i \tag{3}$$

where the noise ε is sampled from the distribution $\mathcal{N}(0,1)$, and cp is computed without noise in order to get stable prototypes. Then, concatenate the latent encoding z_i with the word embedding of the corresponding category and input the concatenated representation into the decoder $\mathcal{F}_{dec}$ to obtain the reconstructed feature x_i':

$$x_i' = \mathcal{F}_{dec}(C(w_i, z_i)) \tag{4}$$

Similar to a standard Variational Autoencoder (VAE) [20], the loss function comprises two components. The first component employs Kullback-Leibler (KL) divergence to measure and constrain the distance between the posterior distribution $q(z|\hat{x}_i)$ and the prior distribution $p(z)$. This ensures that the latent distribution generated by the encoder remains close to the predefined prior distribution, promoting regularization and stability in the latent space:

$$\mathcal{L}_{KL} = \frac{1}{B}\sum_{i=1}^{B} D_{KL}\big(q(z_i \mid \hat{x}_i) \| p(z_i)\big) \tag{5}$$

where B represents the batch size. The mean squared error (MSE) measures the difference between the input x_i and the reconstructed feature x_i', encouraging the decoder to generate reconstructed features that closely match the original input:

$$\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_{i=1}^{B} \text{MSE}(x_i', x_i) \tag{6}$$

where N represents the channel dimension of the input feature x_i. The total loss of MCVA is defined as follows:

$$\mathcal{L}_{\text{MCVA}} = \mathcal{L}_{\text{rec}} + \beta L_{\text{KL}} \tag{7}$$

where β is a hyperparameter. Through the masking and restoration mechanism in the MCVA, the model is compelled to predict the masked elements using the concatenated conditional reference (e.g., word embeddings). When the feature contains category-generalized semantics, the model can accurately predict the masked portions based on the conditional reference. Conversely, when the feature includes biased semantics, the masked elements become difficult to restore. As a result, the semantic bias in the features is suppressed, while category-generalized information is enhanced. Additionally, by leveraging the VAE model to extract latent features from the concatenated input, the generalized semantics are effectively fused into the feature representation, further improving its robustness and discriminative power.

3.4 Task-Aware Feature Aggregation

Traditional meta-learning models typically rely on class prototypes to enhance features for both classification and regression. Ideally, a class prototype should contain only generalized semantics, free from sample-specific variations. However, in object detection tasks, the model must also account for scale information, which is inherently specific to individual samples. To address this, in addition to the MCVA for conditional prototype learning, we introduce the Task-Aware Feature Aggregation (TFA) module. The TFA module is designed to enhance features in different ways for the classification and regression sub-tasks, enhancing the detection performance.

Regression Task Given the absence of scale information in conditional prototypes, we incorporate support image features to assist in the regression task. Specifically, the Contextual Feature Enhancement (CAE) module Fig. 2 takes as input the query features $F_q \in \mathbb{R}^{b\times c\times h_1\times w_1}$ and the support features $F_s \in \mathbb{R}^{n\times c\times h_2\times w_2}$, which are extracted from stages 1–3 of the shared backbone. Here, (h_1,h_2), (w_1,w_2), c and n denote query image and support image height, width, channel number, and number of categories, respectively. The support features are refined through the operation $F_s' = \{M_{pool}(F_s) - A_{pool}(F_s)\} \in \mathbb{R}^{b\times c\times \frac{h_2}{2}\times\frac{w_2}{2}}$, where M_{pool} and A_{pool}, where M_{pool} and A_{pool} represent max pooling and average pooling, respectively. The benefit of this design is that it allows the model to focus on salient features and reduce computational effort, mitigating the risk of overfitting. These refined support features are then used to enhance the query features with the attention matrix $A \in \mathbb{R}^{b\times(h_1w_1)\times(n\times(\frac{h_2w_2}{4}))}$ is computed by:

$$A = \mathrm{softmax}\left(\frac{((F_qW)\cdot(F_s'W)^T)}{\sqrt{c'}}\right) \tag{8}$$

where W represents the FC-layer parameters shared by F_s' and F_q, projecting them from c-dimensional to c'-dimensional space. Then, the enhanced query features $\hat{F}_q \in \mathbb{R}^{b\times c\times h_1\times w_1}$ are obtained using the following approach:

$$\hat{F}_q = F_q + \alpha \times AF_s' \tag{9}$$

where α is a learnable parameter initialized to 0. The $\hat{F}_q$ are fed into the RPN for the training and get proposals P. Then we got the proposal feature $q^{'} = \Omega(Align(\hat{F}_q, P))$, where Ω is backbone network after align, while $Align$ stands for the ROIAlign operation. Finally, proposal feature $q^{'}$ is implemented for the regression training:

$$\mathcal{L}_{reg} = \mathcal{L}_{L1}(\mathcal{F}_{reg}(q'), b) \tag{10}$$

Methods		split 1					split 2					split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW [19]	ICCV19	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
Meta RCNN [42]	ICCV19	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/ cos [35]	ICML20	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [38]	ECCV20	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
DCNet [18]	CVPR21	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
QA-FewDet [12]	ICCV21	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5
FSCE [34]	CVPR21	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
DeFRCN [27]	ICCV21	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4
KFSOD [17]	CVPR22	44.6	45.2	54.4	60.9	65.8	37.8	38.4	43.1	48.1	50.4	34.8	42.7	44.1	52.7	53.9
MRSN [25]	ECCV22	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
Meta FR-CNN [13]	AAAI22	43.0	54.6	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
σ-ADP [6]	ICCV23	52.3	55.5	63.1	65.9	66.7	**42.7**	45.8	48.7	54.8	56.3	47.8	51.8	56.8	60.3	62.4
ICPE [24]	AAAI23	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	**50.9**	53.1	55.3	60.6	60.1
VFA [14]	AAAI23	**57.7**	**64.6**	**64.7**	67.2	67.4	41.4	**46.2**	**51.1**	51.8	51.6	48.9	**54.8**	56.6	59.0	58.9
FPD [37]	AAAI24	48.1	62.2	64.0	67.6	68.4	29.8	43.2	47.7	52.0	53.9	44.9	53.8	**58.1**	61.6	62.9
FM-FSOD [11]	CVPR24	40.1	53.5	57.0	**68.6**	**72.0**	33.1	36.3	48.8	**54.8**	**64.7**	39.2	50.2	55.7	**63.4**	**68.1**
CPL	Ours work	**60.6**	**68.2**	**69.5**	**70.9**	**70.2**	**43.0**	**51.5**	**55.5**	**55.9**	**56.9**	**54.0**	**58.5**	**60.9**	**64.1**	**63.0**

Table 1: Few-shot object detection performance (nAP50) on PASCAL VOC dataset. We evaluate the performance on three different splits. The best and second-best results are colored red and blue.

where $\mathcal{L}_{L1}$ stands for the smoothl1loss, b is the ground truth bounding boxes. This design enhances the scale information of the feature with support features, which contain the sample-specific scale information. The fused feature can achieve better regression performance for object detection.

Classification Task Condional prototypes containing generalized category information can significantly enhance and benefit the classification task. In addition the features after MCVA restored contain richer category information. We utilize the proposal feature q and the restored feature x' from the MCVA to improve sample diversity and enrich feature representations to effectively improve the classification performance. It is worth noting that q and q' are acquired differently. q is an Align operation performed on the original feature F_q (i.e., $q = \Omega(Align(F_q, P))$). The loss function is formulated as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\mathcal{F}_{cls}(q \odot cp_i), c) + \lambda \mathcal{L}_{\mathrm{CE}}(\mathcal{F}_{cls}(x' \odot cp_i), c) \tag{11}$$

where $\mathcal{L}_{cls}$ represents the cross-entropy loss. c represents ground truth category of the object. The conditional prototype cp_i is derived from Eq. 3, $\odot$ denotes channel-wise multiplication.

3.5 Optimization Objective

Our few-shot object detection framework follows a two-stage training paradigm: base class training and novel class fine-tuning. In this work, during the novel class fine-tuning phase, we freeze the backbone network while fine-tuning the remaining components. The training settings for the novel class are the same as the base class training. The overall optimization objective of our model is defined as follows, where $\mathcal{L}_{cls}^{\mathrm{RPN}}$ and $\mathcal{L}_{reg}^{\mathrm{RPN}}$ are the classification and regression losses in RPN:

$$\mathcal{L} = \mathcal{L}_{cls}^{\mathrm{RPN}} + \mathcal{L}_{reg}^{\mathrm{RPN}} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{\mathrm{MCVA}} \tag{12}$$

By jointly training the MCVA and TFA modules, our model effectively refines the semantic bias in prototypes while simultaneously enhancing features for classification and regression tasks through distinct mechanisms, leading to more robust and accurate detection. As a result, the proposed Conditional Prototype Learning (CPL) framework achieves superior performance on the Few-Shot Object Detection (FSOD) task.

Methods		Shot 10	Shot 30
Fine-tuning			
TFA [35]	ICML20	10.0	13.7
MPSR [38]	ECCV20	9.8	14.1
FSCE [34]	CVPR21	11.1	15.3
DeFRCN [27]	ICCV21	**18.5**	**22.6**
KFSOD [47]	CVPR22	18.5	20.3
Meta-learning			
Meta RCNN [42]	ICCV19	8.7	12.4
FsDetView [39]	ECCV20	12.5	14.7
QA-FewDet [12]	ICCV21	11.6	16.5
Meta FR-CNN [13]	AAAI22	12.7	16.6
VFA [14]	AAAI23	16.2	18.9
FPD [37]	AAAI24	16.5	20.1
CPL	Ours work	**17.8**	**21.3**

Table 2: Few-shot detection performance for the novel classes on MS COCO dataset. The upper half is the Fine-tuning method and the lower half is the meta-learning method.

mask_ratio	shot 1	2	3	5	10
0	58.9	64.6	68.2	69	69.3
0.2	59.2	67.3	68.5	69.8	69.1
0.5	**60.6**	**68.2**	**69.5**	70.9	**70.2**
0.9	56.6	64.6	67.9	**71.1**	69.9

Table 3: Analysis of mask ratio in MCVA.

Method	MCVA	TFA	shot 1	3	5
Baseline			40.2	54.0	55.0
Ours	✓		53.0	67.5	69.9
		✓	53.2	65.2	68.3
	✓	✓	**60.6**	**69.5**	**70.9**

Table 4: Ablation study of different components.

Method	Text Encoder	shot 1	2	3	5	10
Ours	w/o reference	54.9	64.3	67.4	69.1	68.9
	CLIP [28]	**60.6**	**68.2**	**69.5**	**70.9**	**70.2**
	Word2Vec [26]	56.4	66.2	67.5	69.6	69.0

Table 5: Analysis of different text encoders.

4 Experiments

4.1 Experimental Setting

Benchmarks. As in previous work, we conduct related experiments on two widely used public object detection datasets, PASCAL VOC [7] and MS COCO [21]. We use the data splits and annotations provided by TFA for fair comparisons.

PASCAL VOC. The dataset contains a total of 20 classes, and the 20 classes are divided into 3 groups, each group consisting of 15 base classes and 5 novel classes. And the VOC2007 and VOC2012 train/val sets are used for training and the VOC2007 test set is used for evaluation. For each novel class set, we have K={1, 2, 3, 5, 10} shots settings. Calculate the Average Precision (IoU=0.5) for the novel classes as the performance evaluation metric.

MS COCO. The dataset contains 80 classes, with the 20 classes intersecting the PASCAL VOC as novel classes and the other 60 classes as base classes. For each novel class set, we have K=10, 30 shots settings. Calculate the Average Precision (IoU=0.5:0.95) for the novel classes as the performance evaluation metric.

Implementation Details. Our approach is realized by MMDetection [4]. We use ResNet-101 [15] pre-trained on ImageNet as the backbone. We resize the query images to a maximum of 1333×800 pixels and adjust the size of cropped instances from the support images to 224×224 pixels. The text encoder uses CLIP-ViT-B/32 [28]. Our model is trained on 2×RTX3090 Nvidia GPUs with a total batch size of 8 using the SGD optimizer. During the base training stage, the hyperparameter λ=0.2 and the model is trained for 30k/165k iterations on the VOC and COCO datasets. The learning rate is set to 0.005 and decays by a factor of 0.1 at 17k/92k iterations. In the fine-tuning stage, the learning rate is set to 0.001 and the λ=0.5 for 1shot settings, λ=0.2 for other settings. Unless otherwise specified, we keep other hyperparameters the same as Meta R-CNN [42].

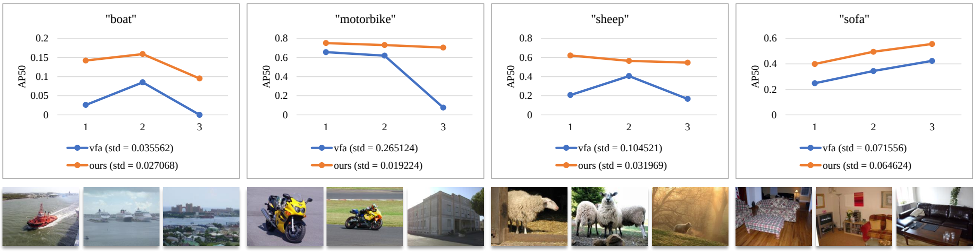

Figure 3: Detection performance varies with different training samples. And conducted training under the 1-shot setting, using the Standard Deviation (std) to measure stability.

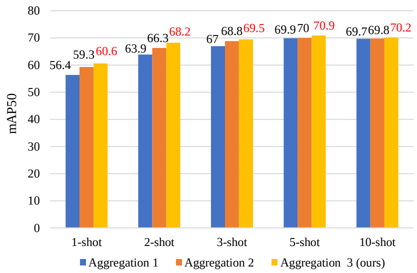

Figure 4: Analysis of different aggregation methods.

λ	Shot 1	2	3	5	10
0.05	54.5	65.7	67.3	70	70
0.2	59.1	**68.2**	**69**	**70.9**	**70.4**
0.5	**60.6**	68.1	68.6	**70.9**	70.2
1	57.9	66	67.7	69.5	68.2

Table 6: Analysis of the hyperparameter λ.

4.2 Main Results

PASCAL VOC. Table 1 shows the experimental results on Pascal VOC, where our method significantly outperforms previous methods and achieves the best (12/15) in most settings. Compared to the previous best results, our approach achieves a maximum improvement of 5.3% (split 2 2-shot), demonstrating its effectiveness.

MS COCO. Table 2 shows the experimental results on COCO. Our method outperforms all meta-learning based methods and outperforms the vast majority of fine-tuning based methods, and is only about 1% lower than DeFRCN. Our methods focuses on the proposed two components and does not use advanced techniques. For example, DeFRCN's gradient decoupling or the transformer based architecture. Also, the COCO dataset is very challenging and we believe that other improvements can be made to further enhance the performance.

4.3 Analyses

We conducted a series of comprehensive experiments on Pascal VOC to validate the effectiveness of the proposed method. If not stated otherwise, the experiments were conducted in the split1 setting by default.

Ablation Study of Different Components. We show the results of different components applied to the baseline Meta RCNN [42] in Table 4. It can be seen that the introduction of MCVA significantly enhances the overall model performance. Building on this, the addition of TFA further achieves the best performance.

Analysis of Different Text Encoders. We implement CLIP [28] to get word embeddings as a semantic reference. In this experiment, we implement Word2Vec [26] for comparison.

The results are shown in Table 5. Both the word embedding methods help the model achieve better performance, while CILP performs better.

Analysis of Sample Sensitive. To validate the influence of different training samples in our model, we choose different samples with a semantic bias for novel classes for 1-shot learning. The results are shown in Fig. 3, where VFA [14] is implemented for comparison. From the results, we can observe that the detection results fluctuate along with different training samples, which shows the influence of sample-specific particularities. Our model achieves more stable performance with different training samples, which demonstrates the effectiveness of our method.

Analysis of Task-aware Feature Aggregation. We evaluate the aggregation strategies of TFA to search for relations between different features and tasks. Specifically, in the experiments, three different aggregation strategies are conducted. Aggregation 1 aggregates the conditional prototype with CAE-enhanced features for classification. Aggregation 2 extends aggregation 1 by further utilizing the aggregated features for regression. Aggregation 3 aggregates the conditional prototype with the original features for classification (same as the model setting). The experimental results are shown in Fig. 4, where the setting of the proposed TFA achieves the best performance. The experiments show the effectiveness of the CAE-enhanced features for regression and prototype features for classification in the detection task.

Analysis of Mask Ratio in MCVA. We conducted the hyper-parameter analysis of λ of Eq. 11, the results are shown in Table 4. It can be found that the model achieves the best performance with the parameter set as 0.5. Additionally, when the 5/10-shot settings are implemented, the difference is not significant. These results show the MCVA can properly reduce the semantic bias, especially on scarce samples. Additionally, when the masking ratio is set to 0.9, the model still achieves similar performance as the low mask rates, which proves the semantic reference can help restoration of features. Thus, the semantic information is properly fused into the model.

Hyperparameter Analysis. In this experiment, we analyze the influence of hyperparameter λ in Eq. 11. The results are shown in Table 6, where the overall performance of the model is best for $\lambda = 0.2$.

5 Conclusion

In this paper, we explore the issue of semantic bias in class prototypes for few-shot object detection (FSOD) under the meta-learning paradigm. To mitigate this problem, we propose MCVA to generate robust conditional prototypes based on conditional references, enhancing the model's generalization ability to novel classes. Additionally, we introduced TFA to optimize the feature representation for the regression and sub-tasks respectively. By combining MCVA and TFA, our CPL strategy achieves state-of-the-art performance. Extensive experiments on PASCAL VOC and MS COCO validate its effectiveness.

Acknowledgement

This work is partially supported by the key project of Humanities and Social Sciences under the Chongqing Ministry of Education (Grant No. 24sKD134), the National Natural Science Foundation of China Youth Program (Grant No. 62306053), the Graduate Innovation Project of Chongqing University of Technology (Grant No. gzlcx20242047).

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, and Dahua Lin. Few-shot object detection via association and discrimination. *Advances in neural information processing systems*, 34:16570–16581, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

[6] Jinhao Du, Shan Zhang, Qiang Chen, Haifeng Le, Yanpeng Sun, Yao Ni, Jian Wang, Bin He, and Jingdong Wang. σ-adaptive decoupled prototype for few-shot object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18904–18914. IEEE, 2023.

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

[8] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, volume 1804, pages 1–6. Springer Berlin/Heidelberg, Germany, 2018.

[9] R Girshick. Fast r-cnn. computer science. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[10] 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.

[11] Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28608–28618, 2024.

[12] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3263–3272, 2021.

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

[14] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via variational feature aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 755–763, 2023.

[15] 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.

[16] 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.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[18] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. 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*, pages 10185–10194, 2021.

[19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8420–8429, 2019.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv.org*, 2014.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] 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.

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[24] Xiaonan Lu, Wenhui Diao, Yongqiang Mao, Junxi Li, Peijin Wang, Xian Sun, and Kun Fu. Breaking immutable: Information-coupled prototype elaboration for few-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1844–1852, 2023.

[25] Tianxue Ma, Mingwei Bi, Jian Zhang, Wang Yuan, Zhizhong Zhang, Yuan Xie, Shouhong Ding, and Lizhuang Ma. Mutually reinforcing structure with proposal contrastive consistency for few-shot object detection. In *European Conference on Computer Vision*, pages 400–416. Springer, 2022.

[26] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

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

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[30] 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.

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

[32] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.

[33] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

[34] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7352–7362, 2021.

[35] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning*, pages 9919–9928. PMLR, 2020.

[36] Yanjie Wang, Xu Zou, Luxin Yan, Sheng Zhong, and Jiahuan Zhou. Snida: Unlocking few-shot object detection with non-linear semantic decoupling augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12544–12553, June 2024.

[37] Zichen Wang, Bo Yang, Haonan Yue, and Zhenghao Ma. Fine-grained prototypes distillation for few-shot object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5859–5866, 2024.

[38] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 456–472. Springer, 2020.

[39] Yang Xiao, Vincent Lepetit, and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3090–3106, 2022.

[40] Honghui Xu, Xinqing Wang, Faming Shao, Baoguo Duan, and Peng Zhang. Few-shot object detection via sample processing. *IEEE Access*, 9:29207–29221, 2021.

[41] Jingyi Xu, Hieu Le, and Dimitris Samaras. Generating features with increased crop-related diversity for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19713–19722, 2023.

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

[43] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=JWOiYxMG92s.

[44] Ze Yang, Chi Zhang, Ruibo Li, Yi Xu, and Guosheng Lin. Efficient few-shot object detection via knowledge inheritance. *IEEE Transactions on Image Processing*, 32: 321–334, 2022.

[45] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12832–12843, 2023.

[46] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[47] Shan Zhang, Lei Wang, Naila Murray, and Piotr Koniusz. Kernelized few-shot object detection with efficient integral aggregation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19207–19216, 2022.

[48] Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13008–13017, 2021.

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.