Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection

Chenchen Zhu Fangyi Chen Uzair Ahmed Zhiqiang Shen Marios Savvides
Carnegie Mellon University
{chenchez, fangyic, uzaira, zhiqians, marioss}@andrew.cmu.edu

Abstract

Few-shot object detection is an imperative and long-lasting problem due to the inherent long-tail distribution of real-world data. Its performance is largely affected by the data scarcity of novel classes. But the semantic relation between the novel classes and the base classes is constant regardless of the data availability. In this work, we investigate utilizing this semantic relation together with the visual information and introduce explicit relation reasoning into the learning of novel object detection. Specifically, we represent each class concept by a semantic embedding learned from a large corpus of text. The detector is trained to project the image representations of objects into this embedding space. We also identify the problems of trivially using the raw embeddings with a heuristic knowledge graph and propose to augment the embeddings with a dynamic relation graph. As a result, our few-shot detector, termed SRR-FSD, is robust and stable to the variation of shots of novel objects. Experiments show that SRR-FSD can achieve competitive results at higher shots, and more importantly, a significantly better performance given both lower explicit and implicit shots. The benchmark protocol with implicit shots removed from the pretrained classification dataset can serve as a more realistic setting for future research.

1. Introduction

Deep learning algorithms usually require a large amount of annotated data to achieve superior performance. To acquire enough annotated data, one common way is by collecting abundant samples from the real world and paying annotators to generate ground-truth labels. However, even if all the data samples are well annotated based on our requirements, we still face the problem of few-shot learning. Because long-tail distribution is an inherent characteristic of the real world, there always exist some rare cases that have just a few samples available, such as rare animals, uncommon road conditions. In other words, we are unable to alleviate the situation of scarce cases by simply spending more money on annotation even big data is accessible.

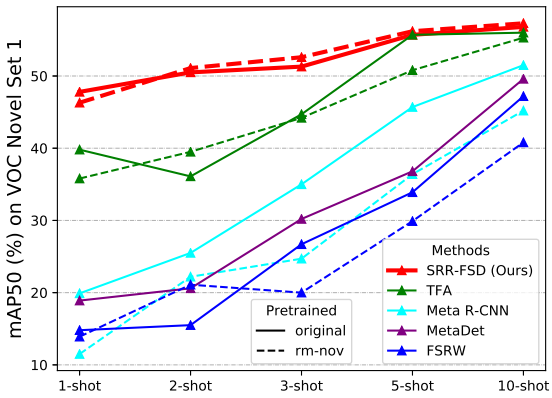

Figure 1. FSOD performance (mAP50) on VOC [13] Novel Set 1 at different shot numbers. Solid line (original) means the pretrained model used for initializing the detector backbone is trained on the original ImageNet [10]. Dashed line (rm-nov) means classes in Novel Set 1 are removed from the ImageNet for the pretrained backbone model. Our SRR-FSD is more stable to the variation of explicit shots (x-axis) and implicit shots (original vs. rm-nov).

Therefore, the study of few-shot learning is an imperative and long-lasting task.

Recently, efforts have been put into the study of few-shot object detection (FSOD) [5, 20, 11, 19, 44, 41, 14, 46, 39, 42, 43]. In FSOD, there are base classes in which sufficient objects are annotated with bounding boxes and novel classes in which very few labeled objects are available. The novel class set does not share common classes with the base class set. The few-shot detectors are expected to learn from limited data in novel classes with the aid of abundant data in base classes and to be able to detect all novel objects in a held-out testing set. To achieve this, most recent few-shot detection methods adopt the ideas from meta-learning and metric learning for few-shot recognition and apply them to conventional detection frameworks, e.g. Faster R-CNN [35], YOLO [34].

Although recent FSOD methods have improved the base-

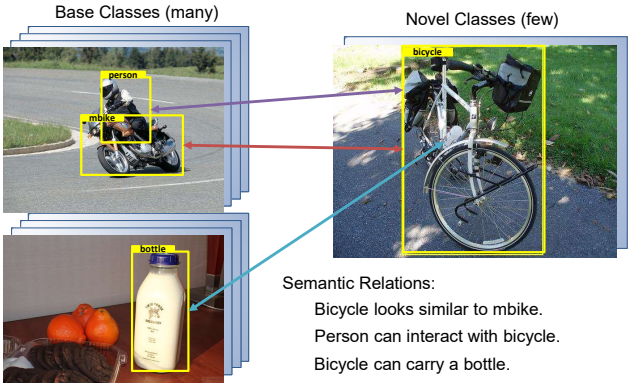

Figure 2. Key insight: the semantic relation between base and novel classes is constant regardless of the data availability of novel classes, which can aid the learning together with visual information.

line considerably, data scarcity is still a bottleneck that hurts the detector's generalization from a few samples. In other words, the performance is very sensitive to the number of both explicit and implicit shots and drops drastically as data becomes limited. The explicit shots refer to the available labeled objects from the novel classes. For example, the 1-shot performance of some FSOD methods is less than half of the 5-shot or 10-shot performance, as shown in Figure 1. In terms of implicit shots, initializing the backbone network with a model pretrained on a large-scale image classification dataset is a common practice for training an object detector. However, the classification dataset contains many implicit shots of object classes overlapped with the novel classes. So the detector can have early access to novel classes and encode their knowledge in the parameters of the backbone. Removing those implicit shots from the pretrained dataset also has a negative impact on the performance as shown in Figure 1. The variation of explicit and implicit shots could potentially lead to system failure when dealing with extreme cases in the real world.

We believe the reason for shot sensitivity is due to exclusive dependence on the visual information. Novel objects are learned through images only and the learning is independent between classes. As a result, visual information becomes limited as image data becomes scarce. However, one thing remains constant regardless of the availability of visual information, i.e. the semantic relation between base and novel classes. For example in Figure 2, if we have the prior knowledge that the novel class "bicycle" looks similar to "motorbike", can have interaction with "person", and can carry a "bottle", it would be easier to learn the concept "bicycle" than solely using a few images. Such explicit relation reasoning is even more crucial when visual information is hard to access [40].

So how can we introduce semantic relation to few-shot detection? In natural language processing, semantic concepts are represented by word embeddings [27, 31] from language models, which have been used in zero-shot learning methods [40, 1]. And explicit relationships are represented by knowledge graphs [28, 4], which are adopted by some zero-shot or few-shot recognition algorithms [40, 30]. However, these techniques are rarely explored in the FSOD task. Also, directly applying them to few-shot detectors leads to non-trivial practical problems, i.e. the domain gap between vision and language, and the heuristic definition of knowledge graph for classes in FSOD datasets (see Section 3.2 and 3.3 for details).

In this work, we explore the semantic relation for FSOD. We propose a Semantic Relation Reasoning Few-Shot Detector (SRR-FSD), which learns novel objects from both the visual information and the semantic relation in an end-to-end style. Specifically, we construct a semantic space using the word embeddings. Guided by the word embeddings of the classes, the detector is trained to project the objects from the visual space to the semantic space and to align their image representations with the corresponding class embeddings. To address the aforementioned problems, we propose to learn a dynamic relation graph driven by the image data instead of pre-defining one based on heuristics. Then the learned graph is used to perform relation reasoning and augment the raw embeddings for reduced domain gap.

With the help of the semantic relation reasoning, our SRR-FSD demonstrates the shot-stable property in two aspects, see the red solid and dashed lines in Figure 1. In the common few-shot settings (solid lines), SRR-FSD achieves competitive performance at higher shots and significantly better performance at lower shots compared to state-of-the-art few-shot detectors. In a more realistic setting (dashed lines) where implicit shots of novel concepts are removed from the classification dataset for the pretrained model, SRR-FSD steadily maintains the performance while some previous methods have results degraded by a large margin due to the loss of implicit shots. We hope the suggested realistic setting can serve as a new benchmark protocol for future research.

We summarize our contributions as follows:

- To our knowledge, our work is the first to investigate semantic relation reasoning for the few-shot detection task and show its potential to improve a strong baseline.
- Our SRR-FSD achieves stable performance w.r.t the shot variation, outperforming state-of-the-art FSOD methods under several existing settings especially when the novel class data is extremely limited.
- We suggest a more realistic FSOD setting in which implicit shots of novel classes are removed from the classification dataset for the pretrained model, and show

that our SRR-FSD can maintain a more steady performance compared to previous methods if using the new pretrained model.

2. Related Work

Object Detection Object detection is a fundamental computer vision task, serving as a necessary step for various down-streaming instance-based understanding. Modern CNN-based detectors can be roughly divided into two categories. One is single-stage detector such as YOLO [34], SSD [26], RetinaNet [24], and FreeAnchor [47] which directly predict the class confidence scores and the bounding box coordinates over a dense grid. The other is multi-stage detector such as Faster R-CNN [35], R-FCN [9], FPN [23], Cascade R-CNN [2], and Libra R-CNN [29] which predict class-agnostic regions of interest and refine those region proposals for one or multiple times. All these methods rely on pre-defined anchor boxes to have an initial estimation of the size and aspect ratio of the objects. Recently, anchor-free detectors eliminate the performance-sensitive hyperparameters for the anchor design. Some of them detect the key points of bounding boxes [22, 48, 12]. Some of them encode and decode the bounding boxes as anchor points and point-to-boundary distances [38, 50, 36, 45, 49]. DETR [3] reformulates object detection as a direct set prediction problem and solve it with transformers. However, these detectors are trained with full supervision where each class has abundant annotated object instances.

Few-Shot Detection Recently, there have been works focusing on solving the detection problem in the limited data scenario. LSTD [5] proposes the transfer knowledge regularization and background depression regularization to promote the knowledge transfer from the source domain to the target domain. [11] proposes to iterate between model training and high-confidence sample selection. RepMet [20] adopts a distance metric learning classifier into the RoI classification head. FSRW [19] and Meta R-CNN [44] predict per-class attentive vectors to reweight the feature maps of the corresponding classes. MetaDet [41] leverages meta-level knowledge about model parameter generation for category-specific components of novel classes. In [14], the similarity between the few shot support set and query set is explored to detect novel objects. Context-Transformer [46] relies on discriminative context clues to reduce object confusion. TFA [39] only fine-tunes the last few layers of the detector. Two very recent papers are MPSR [42] and FSDetView [43]. MPSR develops an auxiliary branch to generate multi-scale positive samples as object pyramids and to refine the prediction at various scales. FSDetView proposes a joint feature embedding module to share the feature from base classes. However, all these methods *depend purely on visual information* and suffer from shot variation.

Semantic Reasoning in Vision Tasks Semantic word embeddings have been used in zero-shot learning tasks to learn a mapping from the visual feature space to the semantic space, such as zero-shot recognition [40] and zero-shot object detection [1, 32]. In [7], semantic embeddings are used as the ground-truth of the encoder TriNet to guide the feature augmentation. In [15], semantic embeddings guide the feature synthesis for unseen classes by perturbing the seen feature with the projected difference between a seen class embedding and a unseen class embedding. In zero-shot or few-shot recognition [40, 30], word embeddings are often combined with knowledge graphs to perform relation reasoning via the graph convolution operation [21]. Knowledge graphs are usually defined based on heuristics from databases of common sense knowledge rules [28, 4]. [8] proposed a knowledge graph based on object co-occurrence for the multi-label recognition task. To our knowledge, the use of word embeddings and knowledge graphs are rarely explored in the FSOD task. Any-Shot Detector (ASD) [33] is the only work that uses word embeddings for the FSOD task. But ASD focuses more on the zero-shot detection and it does not consider the explicit relation reasoning between classes because each word embedding is treated independently.

3. Semantic Relation Reasoning Few-Shot Detector

In this section, we first briefly introduce the preliminaries for few-shot object detection including the problem setup and the general training pipelines. Then based on Faster R-CNN [35], we build our SRR-FSD by integrating semantic relation with the visual information and allowing it to perform relation reasoning in the semantic space. We also discuss the problems of trivially using the raw word embeddings and the predefined knowledge graphs. Finally, we introduce the two-phase training processes. An overview of our SRR-FSD is illustrated in Figure 3.

3.1. FSOD Preliminaries

Conventional object detection problem has a base class set $\mathcal{C}_b$ in which there are many instances, and a base dataset $\mathcal{D}_b$ with abundant images. $\mathcal{D}_b$ consists of a set of annotated images $\{(x_i, y_i)\}$ where x_i is the image and y_i is the annotation of labels from $\mathcal{C}_b$ and bounding boxes for objects in x_i. For few-shot object detection (FSOD) problem, in addition to $\mathcal{C}_b$ and $\mathcal{D}_b$ it also has a novel class set $\mathcal{C}_n$ and a novel dataset $\mathcal{D}_n$, with $\mathcal{C}_b \cap \mathcal{C}_n = \emptyset$. In $\mathcal{D}_n$, objects have labels belong to $\mathcal{C}_n$ and the number of objects for each class is k for k-shot detection. A few-shot detector is expected to learn from $\mathcal{D}_b$ and to quickly generalize to $\mathcal{D}_n$ with a small k such that it can detect all objects in a held-out testing set with object classes in $\mathcal{C}_b \cup \mathcal{C}_n$. We assume all classes in $\mathcal{C}_b \cup \mathcal{C}_n$ have semantically meaningful names so the corresponding semantic embeddings can be retrieved.

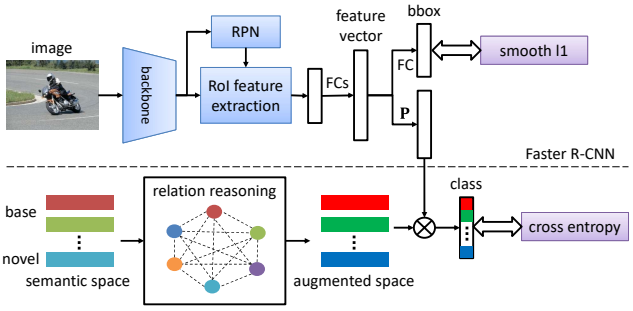

Figure 3. Overview of the SRR-FSD. A semantic space is built from the word embeddings of all corresponding classes in the dataset and is augmented through a relation reasoning module. Visual features are learned to be projected into the augmented space. "⊗": dot product. "FC": fully-connected layer. "**P**": learnable projection matrix.

A typical few-shot detector has two training phases. The first one is the base training phase where the detector is trained on $\mathcal{D}_b$ similarly to conventional object detectors. Then in the second phase, it is further fine-tuned on the union of $\mathcal{D}_b$ and $\mathcal{D}_n$. To avoid the dominance of objects from $\mathcal{D}_b$, a small subset is sampled from $\mathcal{D}_b$ such that the training set is balanced concerning the number of objects per class. As the total number of classes is increased by the size of $\mathcal{C}_n$ in the second phase, more class-specific parameters are inserted in the detector and trained to be responsible for the detection of novel objects. The class-specific parameters are usually in the box classification and localization layers at the very end of the network.

3.2. Semantic Space Projection

Our few-shot detector is built on top of Faster R-CNN [35], a popular two-stage general object detector. In the second-stage of Faster R-CNN, a feature vector is extracted for each region proposal and forwarded to a classification subnet and a regression subnet. In the classification subnet, the feature vector is transformed into a d-dimentional vector $\mathbf{v} \in \mathcal{R}^d$ through fully-connected layers. Then $\mathbf{v}$ is multiplied by a learnable weight matrix $\mathbf{W} \in \mathcal{R}^{N \times d}$ to output a probability distribution as in Eq. (1).

$$\mathbf{p} = softmax(\mathbf{W}\mathbf{v} + \mathbf{b}) \tag{1}$$

where N is the number of classes and $\mathbf{b} \in \mathcal{R}^N$ is a learnable bias vector. Cross-entropy loss is used during training.

To learn objects from both the visual information and the semantic relation, we first construct a semantic space and project the visual feature $\mathbf{v}$ into this semantic space. Specifically, we represent the semantic space using a set of d_e-dimensional word embeddings $\mathbf{W}_e \in \mathcal{R}^{N \times d_e}$ [27] corresponding to the N object classes (including the background class). And the detector is trained to learn a linear projection $\mathbf{P} \in \mathcal{R}^{d_e \times d}$ in the classification subnet (see Figure 3) such that $\mathbf{v}$ is expected to align with its class's word embedding after projection. Mathematically, the prediction of the probability distribution turns into Eq. (2) from Eq. (1).

$$\mathbf{p} = softmax(\mathbf{W}_e\mathbf{P}\mathbf{v} + \mathbf{b}) \tag{2}$$

During training, $\mathbf{W}_e$ is fixed and the learnable variable is $\mathbf{P}$. A benefit is that generalization to novel objects involves no new parameters in $\mathbf{P}$. We can simply expand $\mathbf{W}_e$ with embeddings of novel classes. We still keep the $\mathbf{b}$ to model the category imbalance in the detection dataset.

Domain gap between vision and language. $\mathbf{W}_e$ encodes the knowledge of semantic concepts from natural language. While it is applicable in zero-shot learning, it will introduce the bias of the domain gap between vision and language to the FSOD task. Because unlike zero-shot learning where unseen classes have no support from images, the few-shot detector can rely on both the images and the embeddings to learn the concept of novel objects. When there are very few images to rely on, the knowledge from embeddings can guide the detector towards a decent solution. But when more images are available, the knowledge from embeddings may be misleading due to the domain gap, resulting in a suboptimal solution. Therefore, we need to augment the semantic embeddings to reduce the domain gap. Some previous works like ASD [33] apply a trainable transformation to each word embedding *independently*. But we find leveraging the explicit relationship between classes is more effective for embedding augmentation, leading to the proposal of the dynamic relation graph in Section 3.3.

3.3. Relation Reasoning

The semantic space projection learns to align the concepts from the visual space with the semantic space. But it still treats each class independently and there is no knowledge propagation among classes. Therefore, we further introduce a knowledge graph to model their relationships. The knowledge graph $\mathbf{G}$ is a $N \times N$ adjacency matrix representing the connection strength for every neighboring class pairs. $\mathbf{G}$ is involved in classification via the graph convolution operation [21]. Mathematically, the updated probability prediction is shown in Eq. (3).

$$\mathbf{p} = softmax(\mathbf{G}\mathbf{W}_e\mathbf{P}\mathbf{v} + \mathbf{b}) \tag{3}$$

The heuristic definition of the knowledge graph. In zero-shot or few-shot recognition algorithms, the knowledge graph $\mathbf{G}$ is predefined base on heuristics. It is usually constructed from a database of common sense knowledge rules by sampling a sub-graph through the rule paths such that semantically related classes have strong connections. For example, classes from the ImageNet dataset [10] have a knowledge graph sampled from the WordNet [28]. However, classes in FSOD datasets are not highly semantically

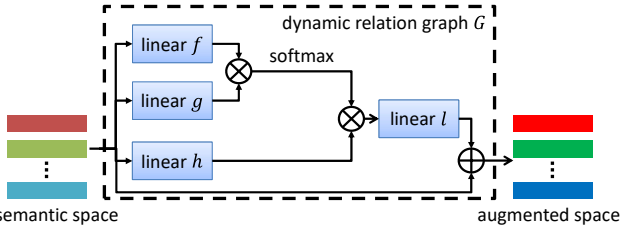

Figure 4. Network architecture of the relation reasoning module for learning the relation graph. "⊗": dot product. "⊕": element-wise plus.

related, nor do they form a hierarchical structure like the ImageNet classes. The only applicable heuristics we found are based on object co-occurrence from [8]. Although the statistics of the co-occurrence are straightforward to compute, the co-occurrence is not necessarily equivalent to semantic relation.

Instead of predefining a knowledge graph based on heuristics, we propose to learn a *dynamic* relation graph driven by the data to model the relation reasoning between classes. The data-driven graph is also responsible for reducing the domain gap between vision and language because it is trained with image inputs. Inspired by the concept of the transformer, we implement the dynamic graph with the self-attention architecture [37] as shown in Figure 4. The original word embeddings $\mathbf{W}_e$ are transformed by three linear layers f, g, h, and a self-attention matrix is computed from the outputs of f, g. The self-attention matrix is multiplied with the output of h followed by another linear layer l. A residual connection [16] adds the output of l with the original $\mathbf{W}_e$. Another advantage of learning a dynamic graph is that it can easily adapt to new coming classes. Because the graph is not fixed and is generated on the fly from the word embeddings. We do not need to redefine a new graph and retrain the detector from the beginning. We can simply insert corresponding embeddings of new classes and fine-tune the detector.

3.4. Decoupled Fine-tuning

In the second fine-tuning phase, we only unfreeze the last few layers of our SRR-FSD similar to TFA [39]. For the classification subnet, we fine-tune the parameters in the relation reasoning module and the projection matrix $\mathbf{P}$. For the localization subnet, it is not dependent on the word embeddings but it shares features with the classification subnet. We find that the learning of localization on novel objects can interfere with the classification subnet via the shared features, leading to many false positives. Decoupling the shared fully-connected layers between the two subnets can effectively make each subnet learn better features for its task. In other words, the classification subnet and the localization subnet have individual fully-connected layers and they are fine-tuned independently.

4. Experiments

4.1. Implementation Details

Our SRR-FSD is implemented based on Faster R-CNN [35] with ResNet-101 [16] and Feature Pyramid Network [23] as the backbone using the MMDetection [6] framework. All models are trained with Stochastic Gradient Descent (SGD) and a batch size of 16. For the word embeddings, we use the L2-normalized 300-dimensional Word2Vec [27] vectors from the language model trained on large unannotated texts like Wikipedia. In the relation reasoning module, we reduce the dimension of word embeddings to 32 which is empirically selected. In the first base training phase, we set the learning rate, the momentum, and the weight decay to 0.02, 0.9, and 0.0001, respectively. In the second fine-tuning phase, we reduce the learning rate to 0.001 unless otherwise mentioned. The input image is sampled by first randomly choosing between the base set and the novel set with a 50% probability and then randomly selecting an image from the chosen set.

4.2. Existing Settings

We follow the existing settings in previous FSOD methods [19, 41, 44, 39] to evaluate our SRR-FSD on the VOC [13] and COCO [25] datasets. For fair comparison and reduced randomness, we use the same data splits and a fixed list of novel samples provided by [19].

VOC The 07 and 12 train/val sets are used for training and the 07 test set is for testing. Out of its 20 object classes, 5 classes are selected as novel and the remaining 15 are base classes, with 3 different base/novel splits. The novel classes each have k annotated objects, where k equals 1, 2, 3, 5, 10. In the first base training phase, our SRR-FSD is trained for 18 epochs with the learning rate multiplied by 0.1 at the 12th and 15th epoch. In the second fine-tuning phase, we train for $500 \times |\mathcal{D}_n|$ steps where $|\mathcal{D}_n|$ is the number of images in the k-shot novel dataset.

We report the mAP50 of the novel classes on VOC with 3 splits in Table 1. In all different base/novel splits, our SRR-FSD achieves a more shot-stable performance. At higher shots like 5-shot and 10-shot, our performance is competitive compared to previous state-of-the-art methods. At more challenging conditions with shots less than 5, our approach can outperform the second-best by a large margin (up to 10+ mAP). Compared to ASD [33] which only reports results of 3-shot and 5-shot in the Novel Set 1, ours is 24.2 and 6.0 better respectively in mAP. We do not include ASD in Table 1 because its paper does not provide the complete results on VOC.

Learning without forgetting is another merit of our SRR-FSD. After generalization to novel objects, the performance

Method / shot	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW [19]	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
MetaDet [41]	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 [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 [39]	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**
SRR-FSD (Ours)	**47.8**	**50.5**	**51.3**	55.2	**56.8**	**32.5**	**35.3**	**39.1**	**40.8**	43.8	**40.1**	**41.5**	**44.3**	46.9	46.4

Table 1. FSOD evaluation on VOC. We report the mAP with IoU threshold 0.5 (mAP50) under 3 different sets of 5 novel classes with a small number of shots.

Shot	Method	Base AP50	Novel AP50
3	Meta R-CNN [44]	64.8	35.0
	TFA [39]	79.1	44.7
	Ours base only	77.7	n/a
	SRR-FSD (Ours)	78.2	51.3
10	Meta R-CNN [44]	67.9	51.5
	TFA [39]	78.4	56.0
	Ours base only	77.7	n/a
	SRR-FSD (Ours)	78.2	56.8

Table 2. FSOD performance for the base and novel classes on Novel Set 1 of VOC. Our SRR-FSD has the merit of learning without forgetting.

Shot	Method	AP	AP50	AP75
10	FSRW [19]	5.6	12.3	4.6
	MetaDet [41]	7.1	14.6	6.1
	Meta R-CNN [44]	8.7	19.1	6.6
	TFA [39]	10.0	-	9.3
	MPSR [42]	9.8	17.9	9.7
	SRR-FSD (Ours)	**11.3**	**23.0**	**9.8**
30	FSRW [19]	9.1	19.0	7.6
	MetaDet [41]	11.3	21.7	8.1
	Meta R-CNN [44]	12.4	25.3	10.8
	TFA [39]	13.7	-	13.4
	MPSR [42]	14.1	25.4	**14.2**
	SRR-FSD (Ours)	**14.7**	**29.2**	13.5

Table 3. FSOD performance of the novel classes on COCO.

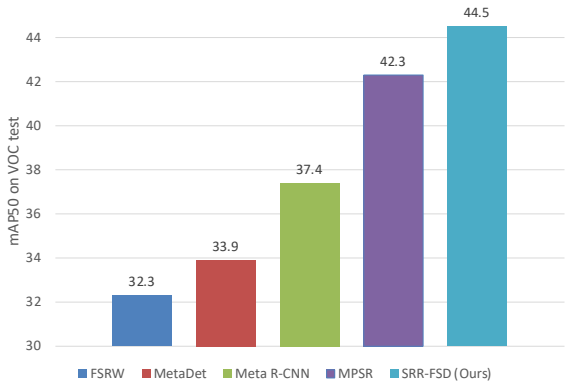

Figure 5. 10-shot cross domain performance on the 20 novel classes under COCO to VOC.

on the base objects does not drop at all as shown in Table 2. Both base AP and novel AP of our SRR-FSD compare favorably to previous methods based on the same Faster R-CNN with ResNet-101. The base AP even increases a bit probably due to the semantic relation reasoning from limited novel objects to base objects.

COCO The `minival` set with 5000 images is used for testing and the rest images in train/val sets are for training. Out of the 80 classes, 20 of them overlapped with VOC are the novel classes with $k = 10, 30$ shots per class and the remaining 60 classes are base. We train the SRR-FSD on the base dataset for 12 epochs using the same setting as MMDetection [6] and fine-tune it for a fixed number of $10 \times |\mathcal{D}_b|$ steps where $|\mathcal{D}_b|$ is the number of images in the base dataset. Unlike VOC, the base dataset in COCO contains unlabeled novel objects, so the region proposal network (RPN) treats them as the background. To avoid omitting novel objects in the fine-tuning phase, we unfreeze the RPN and the following layers. Table 3 presents the COCO-style averaged AP. Again we consistently outperform previous methods including FSRW [19], MetaDet [41], Meta R-CNN [44], TFA [39], and MPSR [42].

COCO to VOC For the cross-domain FSOD setting, we follow [19, 41] to use the same base dataset with 60 classes as in the previous COCO within-domain setting. The novel dataset consists of 10 samples for each of the 20 classes from the VOC dataset. The learning schedule is the same as the previous COCO within-domain setting except the learning rate is 0.005. Figure 5 shows that our SRR-FSD achieves the best performance with a healthy 44.5 mAP, indicating better generalization ability in cross-domain situations.

4.3. A More Realistic Setting

The training of the few-shot detector usually involves initializing the backbone network with a model pretrained on large-scale object classification datasets such as ImageNet [10]. The set of object classes in ImageNet, i.e. $\mathcal{C}_0$, is highly overlapped with the novel class set $\mathcal{C}_n$ in the existing settings. This means that the pretrained model can get early access to large amounts of object samples, i.e. *implicit*

	Novel Set 1					Novel Set 2					Novel Set 3				
Method / shot	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW [19]	13.9	21.1	20.0	29.9	40.8	13.5	14.2	20.6	20.7	36.8	16.2	22.2	26.8	37.0	41.5
Meta R-CNN [44]	11.5	22.2	24.7	36.4	45.2	10.1	16.9	22.7	29.6	40.1	10.0	21.7	27.1	32.8	41.6
TFA [39]	35.8	39.5	44.2	50.8	55.3	18.8	26.0	33.2	31.3	39.2	25.6	32.6	36.4	**43.7**	**48.5**
SRR-FSD (Ours)	**46.3**	**51.1**	**52.6**	**56.2**	**57.3**	**31.0**	**29.9**	**34.7**	**37.3**	**41.7**	**39.2**	**40.5**	**39.7**	42.2	45.2

Table 4. FSOD performance (mAP50) on VOC under a more realistic setting where novel classes are removed from the pretrained classification dataset to guarantee $\mathcal{C}_0 \cap \mathcal{C}_n = \emptyset$. Our SRR-FSD is more robust to the loss of implicit shots comparing with Table 1.

shots, from novel classes and encode their knowledge in the parameters before it is further trained for the detection task. Even the pretrained model is optimized for the recognition task, the extracted features still have a big impact on the detection of novel objects (see Figure 1). However, some rare classes may have highly limited or valuable data in the real world that pretraining a classification network on it is not realistic.

Therefore, we suggest a more realistic setting for FSOD, which extends the existing settings. In addition to $\mathcal{C}_b \cap \mathcal{C}_n = \emptyset$, we also require that $\mathcal{C}_0 \cap \mathcal{C}_n = \emptyset$. To achieve this, we systematically and hierarchically remove novel classes from $\mathcal{C}_0$. For each class in $\mathcal{C}_n$, we find its corresponding synset in ImageNet and obtain its full hyponym (the synset of the whole subtree starting from that synset) using the ImageNet API [1]. The images of this synset and its full hyponym are removed from the pretrained dataset. And the classification model is trained on a dataset with no novel objects. We provide the list of WordNet IDs for each novel class to be removed in the supplementary materials.

We notice that CoAE [18] also proposed to remove all COCO-related ImageNet classes to ensure the model does not "foresee" the unseen classes. As a result, a total of 275 classes are removed from ImageNet including both the base and novel classes in VOC [13], which correspond to more than 300k images. We think the loss of this much data may lead to a worse pretrained model in general. So the pretrained model may not be able to extract features strong enough for down-streaming vision tasks compared with the model trained on full ImageNet. Our setting, on the other hand, tries to alleviate this effect as much as possible by only removing the novel classes in VOC Novel Set 1, 2, and 3 respectively, which correspond to an average of 50 classes from ImageNet.

Under the new realistic setting, we re-evaluate previous methods using their official source code and report the performance on the VOC dataset in Table 4. Our SRR-FSD demonstrates superior performance to other methods under most conditions, especially at challenging lower shot scenarios. More importantly, our SRR-FSD is less affected by the loss of implicit shots. Compared with results in Table 1, our performance is more stably maintained when novel objects are only available in the novel dataset.

[1] http://image-net.org/download-API

4.4. Ablation Study

In this section, we study the contribution of each component. Experiments are conducted on the VOC dataset. Our baseline is the Faster R-CNN [35] with ResNet-101 [16] and FPN [23]. We gradually apply the Semantic Space Projection (SSP 3.2), Relation Reasoning (RR 3.3) and Decoupled Fine-tuning (DF 3.4) to the baseline and report the performance in Table 5. We also compare three different ways of augmenting the raw word embeddings in Table 6, including the trainable transformation from ASD [33], the heuristic knowledge graph from [8], and the dynamic graph from our proposed relation reasoning module.

Semantic space projection guides shot-stable learning. The baseline Faster R-CNN can already achieve satisfying results at 5-shot and 10-shot. But at 1-shot and 2-shot, performance starts to fall apart due to exclusive dependence on images. The semantic space projection, on the other hand, makes the learning more stable to the variation of shot numbers (see 1st and 2nd entries in Table 5). The space projection guided by the semantic embeddings is learned well enough in the base training phase so it can be quickly adapted to novel classes with a few instances. We can observe a major boost at lower shot conditions compared to baseline, i.e. 7.9 mAP and 2.4 mAP gain at 1-shot and 2-shot respectively. However, the raw semantic embeddings limit the performance at higher shot conditions. The performance at 5-shot and 10-shot drops below the baseline. This verifies our argument about the domain gap between vision and language. At lower shots, there is not much visual information to rely on so the language information can guide the detector to a decent solution. But when more images are available, the visual information becomes more precise then the language information starts to be misleading. Therefore, we propose to refine the word embeddings for a reduced domain gap.

Relation reasoning promotes adaptive knowledge propagation. The relation reasoning module explicitly learns a relation graph that builds direct connections between base classes and novel classes. So the detector can learn the novel objects using the knowledge of base objects besides the visual information. Additionally, the relation

	Components			Shots in Novel Set 1				
	SSP	RR	DF	1	2	3	5	10
Faster R-CNN [35]				32.6	44.4	46.3	49.6	55.6
	✓			40.5	46.8	46.5	47.1	52.2
	✓	✓		44.1	46.0	47.8	51.7	54.7
SRR-FSD	✓	✓	✓	47.8	50.5	51.3	55.2	56.8

Table 5. Ablative performance (mAP50) on the VOC Novel Set 1 by gradually applying the proposed components to the baseline Faster R-CNN. **SSP**: semantic space projection. **RR**: relation reasoning. **DF**: decoupled fine-tuning.

	Shots in Novel Set 1				
	1	2	3	5	10
+SSP	40.5	46.8	46.5	47.1	52.2
+SSP +TT [33]	39.3	45.7	43.9	49.4	52.4
+SSP +HKG [8]	41.6	45.5	47.8	49.7	52.5
+SSP +RR	44.1	46.0	47.8	51.7	54.7

Table 6. Comparison of three ways of refining the word embeddings, including the trainable transformation from ASD [33], the heuristic knowledge graph from [8], and the dynamic relation graph from our relation reasoning module. **SSP**: semantic space projection. **RR**: relation reasoning. **TT**: trainable transformation. **HKG**: heuristic knowledge graph.

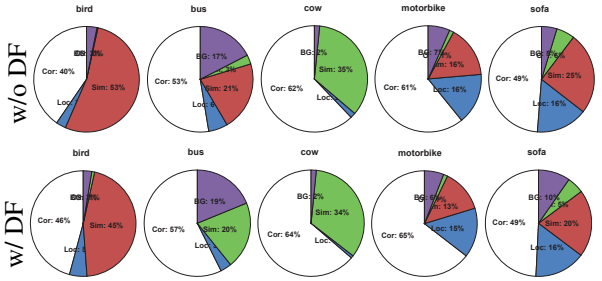

Figure 6. Error analysis of false positives in VOC Novel Set 1 with and without decouple fine-tuning (DF). Detectors are trained with 3 shots. Pie charts indicate the fraction of correct detections (Cor) and top-ranked false positives that are due to poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), or confusion with background or unlabeled objects (BG).

reasoning module also functions as a refinement to the raw word embeddings with a data-driven relation graph. Since the relation graph is updated with image inputs, the refinement tends to adapt the word embeddings for the vision domain. Results in Table 5 (2nd and 3rd entries) confirm that applying relation reasoning improves the detection accuracy of novel objects under different shot conditions. We also compare it with two other ways of refining the raw word embeddings in Table 6. One is the trainable transformation (TT) from ASD [33] where word embeddings are updated with a trainable metric and a word vocabulary. Note that this transformation is applied to each embedding independently which does not consider the explicit relationships between them. The other one is the heuristic knowledge graph (HKG) defined based on the co-occurrence of objects from [8]. It turns out both the trainable transformation and the predefined heuristic knowledge graph are not as effective as the dynamic relation graph in the relation reasoning module. The effect of the trainable transformation is similar to unfreezing more parameters of the last few layers during fine-tuning as shown in the supplementary materials, which leads to overfitting when the shot is low. And the predefined knowledge graph is fixed during training thus cannot be adaptive to the inputs. In other words, the dynamic relation graph is better because it can not only perform explicit relation reasoning but also augment the raw embeddings for reduced domain gap between vision and language.

Decoupled fine-tuning reduces false positives. We analyze the false positives generated by our SRR-FSD with and without decoupled fine-tuning (DF) using the detector diagnosing tool [17]. The effect of DF on reducing the false positives in novel classes is visualized in Figure 6. It shows that most of the false positives are due to misclassification into similar categories. With DF, the classification subnet can be trained independently from the localization subnet to learn better features specifically for classification.

5. Conclusion

In this work, we propose semantic relation reasoning for few-shot object detection. The key insight is to explicitly integrate semantic relation between base and novel classes with the available visual information, which can help to learn the novel concepts better especially when the novel class data is extremely limited. We apply the semantic relation reasoning to the standard two-stage Faster R-CNN and demonstrate robust few-shot performance against the variation of shot numbers. Compared to previous methods, our approach achieves state-of-the-art results on several few-shot detection settings, as well as a more realistic setting where novel concepts encoded in the pretrained backbone model are eliminated. We hope this realistic setting can be a better evaluation protocol for future few-shot detectors. Last but not least, the key components of our approach, i.e. semantic space projection and relation reasoning, can be straightly applied to the classification subnet of other few-shot detectors.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2, 3

[2] 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. 3

[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. 3

[4] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *Aaai*, volume 5. Atlanta, 2010. 2, 3

[5] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 3

[6] 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, 6

[7] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28(9):4594–4605, 2019. 3

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 3, 5, 7, 8

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 3

[10] 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. 1, 4, 6

[11] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1641–1654, 2018. 1, 3

[12] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Object detection with keypoint triplets. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[13] 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(2):303–338, 2010. 1, 5, 7

[14] 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. 1, 3

[15] Jiechao Guan, Zhiwu Lu, Tao Xiang, Aoxue Li, An Zhao, and Ji-Rong Wen. Zero and few shot learning with semantic feature synthesis and competitive learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[16] 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. 5, 7

[17] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 8

[18] 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. 7

[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 International Conference on Computer Vision*, pages 8420–8429, 2019. 1, 3, 5, 6, 7

[20] 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. 1, 3

[21] T.N. Kipf and M. Welling. Semi-supervised classification with graph convolutional network. In *International Conference on Learning Representations (ICLR)*, 2017. 3, 4

[22] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 3

[23] 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. 3, 5, 7

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[26] 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. 3

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and

phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2, 4, 5

[28] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2, 3, 4

[29] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019. 3

[30] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 441–449, 2019. 2, 3

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[32] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, pages 11932–11939, 2020. 3

[33] Shafin Rahman, Salman Khan, Nick Barnes, and Fahad Shahbaz Khan. Any-shot object detection. *arXiv preprint arXiv:2003.07003*, 2020. 3, 4, 5, 7, 8

[34] 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. 1, 3

[35] 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. 1, 3, 4, 5, 7, 8

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[37] 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. 5

[38] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019. 3

[39] 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. 1, 3, 5, 6, 7

[40] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 2, 3

[41] 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. 1, 3, 5, 6

[42] 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*. Springer, 2020. 1, 3, 6

[43] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European conference on computer vision*. Springer, 2020. 1, 3

[44] 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. 1, 3, 5, 6, 7

[45] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[46] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *AAAI*, pages 12653–12660, 2020. 1, 3

[47] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *Advances in neural information processing systems*, 2019. 3

[48] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019. 3

[49] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[50] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3