SNIDA: Unlocking Few-Shot Object Detection with Non-linear Semantic Decoupling Augmentation

Yanjie Wang[1,2], Xu Zou[1,2,*], Luxin Yan[1,2], Sheng Zhong[1,2], Jiahuan Zhou [3]
[1]Huazhong University of Science and Technology, Wuhan 430074, China
[2]National Key Laboratory of Multispectral Information Intelligent Processing Technology, China
[3]Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China
{aiawyj, zoux, yanluxin, zhongsheng}@hust.edu.cn, jiahuanzhou@pku.edu.cn

Abstract

Once only a few-shot annotated samples are available, the performance of learning-based object detection would be heavily dropped. Many few-shot object detection (FSOD) methods have been proposed to tackle this issue by adopting image-level augmentations in linear manners. Nevertheless, those handcrafted enhancements often suffer from limited diversity and lack of semantic awareness, resulting in unsatisfactory performance. To this end, we propose a Semantic-guided Non-linear Instance-level Data Augmentation method (SNIDA) for FSOD by decoupling the foreground and background to increase their diversities respectively. We design a semantic awareness enhancement strategy to separate objects from backgrounds. Concretely, masks of instances are extracted by an unsupervised semantic segmentation module. Then the diversity of samples would be improved by fusing instances into different backgrounds. Considering the shortcomings of augmenting images in a limited transformation space of existing traditional data augmentation methods, we introduce an object reconstruction enhancement module. The aim of this module is to generate sufficient diversity and non-linear training data at the instance level through a semantic-guided masked autoencoder. In this way, the potential of data can be fully exploited in various object detection scenarios. Extensive experiments on PASCAL VOC and MS-COCO demonstrate that the proposed method outperforms baselines by a large margin and achieves new state-of-the-art results under different shot settings.

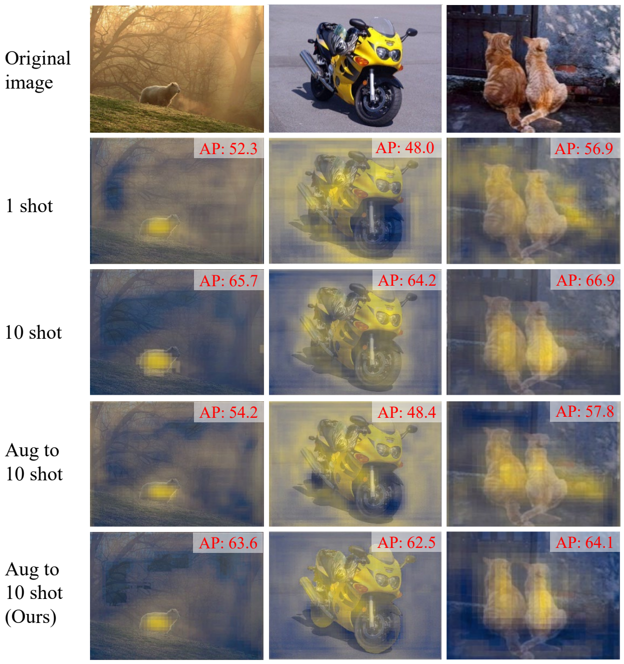

Figure 1. Confidence weighting maps under different shot settings on PASCAL VOC. N-shot means each category with N annotations during finetuning. Aug to 10 shot indicates data augmentation from 1-shot to 10-shot with traditional linear augmentation methods (including flipping, rotating, scaling, and cutting). Our method (row.4 vs row.5) better adapts to the shape characteristics of various categories, which means our method possesses superior semantic awareness capabilities.

1. Introduction

In recent years, deep learning based methods have achieved impressive performance in a variety of visual tasks, such as object recognition [38, 39], and image segmentation [17, 30]. However, it relies heavily on abundant annotated data, which limits their applicability to some realistic scenes, where samples are hard to collect or annotations are expensive. In contrast, humans can quickly grasp a novel concept with few samples. To bridge the gap between the performance of deep learning and the ability of humans, *few-shot Object detection* (FSOD) [9, 34, 45] has attracted much attention. FSOD, typically with only K instances per

[*] *Corresponding author.*

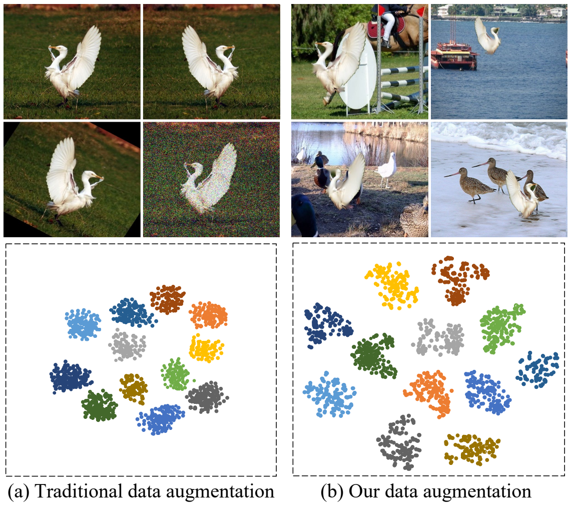

Figure 2. Augmented images and the t-SNE [42] visualization of object proposal embedding learned with (a) traditional data augmentation and (b) our proposed method respectively. Instances from each category form tight clusters, indicating their similarities. In comparison, clusters generated by our method are much more divergent than ones by the traditional linear method. That is, samples augmented by our method would have much greater diversities since they are reconstructed via learned discriminative features (resembling a series of local semantic blurs qualitatively), resulting in a better FSOD performance.

class, easily lacks generalization capability and incurs overfitting since only extremely few annotated instances of previously unseen categories are available for training.

An obvious solution for FSOD is data augmentation [24, 25], which tries to increase the diversity of the input information by directly expanding the number of training labeled instances or indirectly augmenting the features. Most existing data augmentation methods [4, 5, 22, 23, 26, 41, 48] obtain more samples through handcrafted enhancements (e.g. flipping, rotating, scaling, cutting) at levels of images, proposals, or features. However, the common characteristic of these methods is that linear-based augmentations lack semantic awareness and can not generate various training samples with sufficient diversity, since they are not foreground/background sensitive.

As shown in Figure 1, for a ResNet-101 [16] based object detection model, the confidence weighting maps would gradually become salient once more training instances for novel classes are available. That is when there is only one instance, information distinguishing foreground and background tends to be noisy and local. Compared with 1-shot, 10-shot could help converge to salient features and achieve better separation from the background (Figure 1, row.2 and row.3). At present, there is no practical solution to obtain semantic information for few-shot instances. For example, though augmenting 1-shot to 10-shot with traditional methods, such as flipping/rotating/scaling, increases the sample quantity, the performance improvement is limited since the semantic information is not well enhanced and the foreground/background still can not be readily separated (Figure 1, row.3 vs row.4), resulting more susceptible to overfitting due to the lack of diversity. However, Our method (Figure 1, row.4 vs row.5) better adapts to the shape of various categories, which means our method possesses superior semantic awareness capabilities.

In this paper, a novel data augmentation method is proposed to address these issues for FSOD, which decouples the foreground and background to increase their diversity and augments various training data in semantic-guided nonlinear transformation spaces. It consists of the semantic awareness enhancement strategy (SAES) and the object reconstruction enhancement module (OREM) to generate augmented training data at the instance level. The SAES is adopted to decouple foreground and background. Then the diversity of samples would be improved by fusing instances into different backgrounds, so as to further enhance the semantic awareness of the object's discriminative features. To avoid expanding images in a limited linear transformation space, inspired by the powerful nonlinear fitting and reconstruction capabilities of the masked image modeling [18], we introduce OREM to generate diverse data for few-shot samples in a non-linear manner. It randomly masks instances of the novel classes and feeds them into a self-supervised encoder-decoder (e.g. MAE [18]) to export reconstructed images. While the higher Masking Ratio of MAE can indeed introduce more sample diversity, increased randomness, and uncertainty also bring uncontrollable semantic variations, resulting in weaker classification performance. To maintain the high-level semantics of the reconstructed samples, we adopt a language embedding model to guide the high-level knowledge of the MAE's encoder to generate controlled semantic samples.

We have also conducted some pre-experiments. As illustrated in Figure 2, augmented images and t-SNE [42] visualization shows the proposal embedding of augmenting randomly selected 200 PASCAL VOC images to 2400 images by traditional and proposed methods respectively. Instances from each category form tight clusters, indicating their similarities. In comparison, clusters generated by our method are much more divergent than ones by the traditional linear method. That is, samples augmented by our method would have much greater diversities, resulting in a better FSOD performance. To summarize, our contributions are 3-fold:

(1) In this paper, we present a novel data augmentation method for FSOD, which decouples the foreground/background of the novel class objects and increases their diversity in a semantic-guided non-linear manner.

(2) The core idea of our method is non-linear semantic decoupling. Thus, the SAES is adopted to decouple foreground/background and the OREM is introduced to gener-

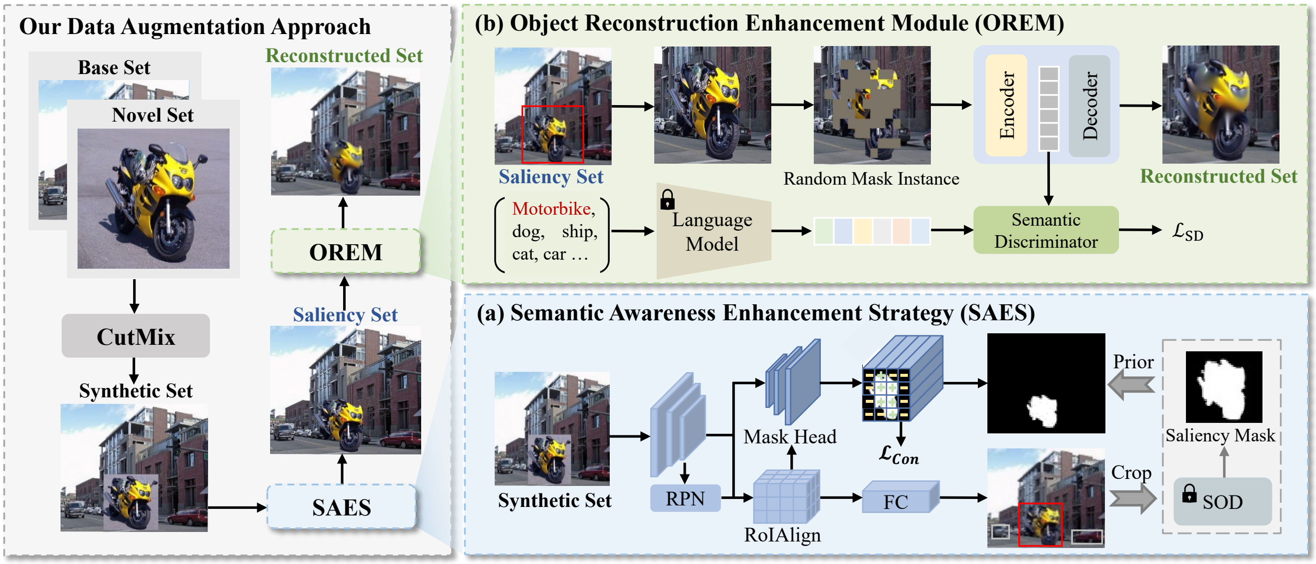

Figure 3. The overall framework. To enhance the generalization ability, we present a novel data augmentation method consisting of two steps to generate diverse data. (a) To decouple the foreground/background of few-shot instances, Semantic Awareness Enhancement Strategy (SAES) is adopted to extract the mask of instances. Then the diversity of samples would be improved by fusing instances into different backgrounds. (b) Object Reconstruction Enhancement Module (OREM) is employed to randomly mask instances and feed them into a semantic-guided masked autoencoder to export the enhanced image in a non-linear manner.

ate diverse instances with high-level semantic invariance.

(3) The proposed method can be readily plugged into existing finetuning-based FSOD methods for further performance improvement. Extensive experiments on PASCAL VOC and MS-COCO with the proposed method outperform baselines by a large margin and achieve new state-of-the-art results under different shot settings.

2. Related Work

2.1. Few-Shot Object Detection

Differing from general object detection [10, 11, 37–39] with adequate annotations, few-shot object detection (FSOD) [3, 7, 21, 46, 54] aims to detect objects with few labeled samples, which could be roughly categorized into the meta-learning fashion and finetuning fashion. For the meta-learning fashion, methods [20, 50] learn-to-learn solving a set of unrelated tasks. The aim is to perform an exemplar search at the instance level using only a few annotated image support sets. FSRW [20] leverages a reweighting module with adequate labeled base classes based on YOLOv2 [36] and transfers to novel classes quickly. FS-DetView [50] introduces a joint feature embedding to make the most of rich feature information originating from base class sharing. These methods usually suffer from the design of complex episodic training schemes. Recently, several finetuning methods [34, 45] attracted more attention compared to meta-learning methods. Two-stage Finetuning Approach [45] (TFA) is the baseline method of finetuning fashion, which trains on fully labeled base classes in the first stage, and finetunes on the balanced support set in the second stage. A simple yet effective work, DeFRCN [34], decouples conflict tasks between class-agnostic RPN and class-relevant RCNN. MFD [49] is designed to explicitly learn three types of commonalities between base classes and novel classes, which are extracted from a memory bank during the fine-tuning phase. However, the overfitting problem easily arises in this fashion with limited instances of novel classes. In this work, we design a novel non-linear instance-level data augmentation method to alleviate the problem.

2.2. Data augmentation in FSOD

Data augmentation is a common tool for improving performance in computer vision, especially in cases where training data is not abundant, such as FSOD. In the recent literature, MPSR [48] and FSOD-SR [23] generate multi-scale positive instances as object pyramids to solve the problem of scale variations. LVC [22] introduces a pseudo-labeling method to source high-quality pseudo-annotations in novel categories. FSCE [41] regards the proposals of different IoU as the intra-image augmentation used in contrastive methods. However, the above linear augmentation methods enhance sample quantity in the limited transformation space, and cannot generate sufficient diversity of data.

NP-RepMet [53] incorporates the negative proposals discarded into the model training, which results in a more robust embedding space. To address the lack of variability in the training data, Halluc [57] introduces a hallucinator network to transfer the shared within-class variation from base classes to novel classes. And TIP [26] introduces transformed guidance consistency loss on predictions from various transformed images. However, most of the above methods rely on image-level augmentation and lack of semantic awareness, resulting in unsatisfactory performance. Different from previous methods for FSOD, our method improves the diversity of samples in a semantic-guided non-linear manner at the instance level.

3. Method

3.1. Problem Setting

Referring to the problem setup [20, 45, 50] used in previous research, we split the general object detection datasets into FSOD datasets. Specifically, object classes are divided into the base class set C_{base} with abundant annotations and the novel class set C_{novel} with only k labeled samples in each class, where C_{base} from the base class dataset d_{base}, C_{novel} from the novel class dataset d_{novel}. There is no intersection between two category sets, $C_{base} \cap C_{novel} = \varnothing$. Our approach follows the finetuning-based methods [34, 45] of FSOD, which can be divided into the base training stage and the novel class finetuning stage.

3.2. Overview

The overall proposed framework is shown in Figure 3. We present a Semantic-guided Non-linear Instance-level Data Augmentation method (SNIDA) for FSOD. Our approach first applies CutMix data augmentation [55] to produce a synthetic set (d_{syn}), which crops entire object patches (P_n) from C_{novel} images and randomly pastes them onto C_{base} images. Next, we adopt the Semantic Awareness Enhancement Strategy to generate a saliency set (d_{sal}), which obtains semantic awareness through foreground/background decoupling and fusing instances into different backgrounds. To further increase the diversity of instances, we employ the Object Reconstruction Enhancement Module to generate the reconstructed set (d_{rec}) during finetuning, which introduces semantic-guided masked image modeling to generate reconstructed images of few-shot instances. The generated d_{syn}, d_{sal}, and d_{rec} are all employed during the finetuning stage.

3.3. Semantic Awareness Enhancement Strategy

To increase the number of novel class samples, we perform CutMix to combine base class images with novel class objects, which crops entire object patches (P_n) from images in d_{novel} and applies simple random data augmentations such as scaling/ flipping/color degradation. The augmented object patches are then pasted onto the images I_b from d_{base}. By repeating this process, we generate synthetic set (d_{syn}) containing samples from both d_{novel} and d_{base}. The visualization of d_{syn} is shown in Figure 4 (col.1).

However, in image-level augmentation methods (CutMix), the discontinuities at the merged boundaries may lead the network to mistakenly treat fixed background features as category-discriminative features, particularly when only a few samples are available. To address this issue, we adopt the Semantic Awareness Enhancement Strategy (SAES), to guide the network by decoupling the foreground and background to extract object semantic discriminative features during finetuning. In our approach, we add a split branch (Mask Head) to our baselines to learn the mask of C_{novel} samples. The input to the network is synthetic set d_{syn} generated by CutMix, and the output is the detection results of all objects as well as the mask of novel class samples.

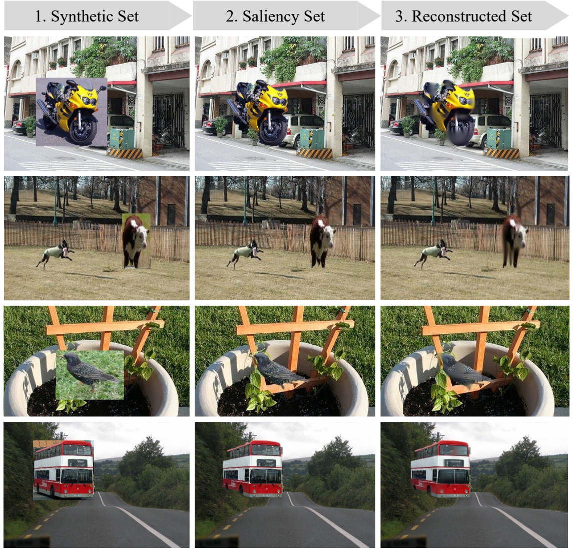

Figure 4. Visualizations of the synthetic, saliency, and reconstructed set on PASCAL VOC. Though the reconstructed set seems to just introduce some local blurs compared to the saliency set, these local blurs are generated via learned discriminative features which could effectively enhance the sample semantic diversity.

As there are no segmentation masks available for few-shot objects, we need to rely on unsupervised methods. Nevertheless, learning dense semantic representations of few-shot objects in an unsupervised setting is a challenging task. Particularly with a low-data regime, the network may prioritize low-level image features over image semantics. Inspired by the unsupervised semantic segmentation method [43], we further introduce the unsupervised saliency object detection (SOD) [33] to learn the mask regions of novel class objects. These regions serve as priors for guiding the network to learn the pixel representation for segmentation. SOD provides only the mask area of salient objects. Therefore, we crop the region (R_n) of novel class objects from d_{syn} as input of SOD. Then, the object saliency mask M_n obtained from SOD is pasted onto the background mask to obtain the mask $M_{I(s)}$ of the input image I as a prior.

Although the mask $M_{I(s)}$ provides some object information, it is not a semantic-level supervision. To address this problem, we introduce a pixel-level contrastive loss to learn the representation of foreground/background, comprising distances between embedding vectors for positive sample pairs and negative sample pairs, as follows:

$$\mathcal{L}_{Con} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} d(v_i, v_j) - \frac{1}{|\mathcal{N}|}\sum_{(i,k)\in\mathcal{N}} d(v_i, v_k) \quad (1)$$

Where, $d(\cdot,\cdot)$ denotes Euclidean distance for measuring the distance between pixel embedding vectors. $\mathbf{v}_i$ denotes the embedding vector of the i-th pixel. $\mathcal{P}$ and $\mathcal{N}$ represent the set of positive and negative sample pairs. The optimization process of this loss is designed to minimize the distance between embedding vectors belonging to the same classes while pushing foreground and background apart. In this way, a pixel embedding space is introduced as a dense semantic representation to get the mask M_I of C_{novel}.

Finally, we assemble the segmented novel class instances into the original base class image I_b as follows:

$$I_{sal} = I_{syn} \odot M_I + I_b \odot (1 - M_I) \quad (2)$$

where $\odot$ is the dot product. By splicing the novel class instance into the image of I_b, we get the I_{sal} image of d_{sal} and the corresponding mask M_I. As shown in Figure 4 (col.2), d_{sal} achieves instance-level augmentation by fusing instances into different backgrounds.

3.4. Object Reconstruction Enhancement Module

In this section, we introduce a semantic-guided nonlinear data augmentation method at the instance level, namely the Object Reconstruction Enhancement Module (OREM), which effectively enhances the diversity of C_{novel} instances during finetuning. We draw inspiration from masked image modeling, such as Masked Autoencoder (MAE) [18] leverage a self-supervised masked encoder to generate the reconstructed image of the input mask tokens. Thus, we utilize the non-linear fitting capability of the MAE to further augment instance diversity. To maximize the impact of MAE on the object, we crop each novel class object from I and its corresponding mask M to obtain the cropped k_{th} novel class objects image I_n^k and mask M_n^k.

The diversity of samples has been improved by fusing instances into different backgrounds in SAES. Thus we expect that part of the foreground area will also be enhanced while the background area remains unchanged in OREM, so as to increase the diversity of the instance. Specifically, we divide the I_n^k image into N non-overlapping patches ($\{x_i^p\}_{i=1}^N$) and the positions of patches are denoted as $P \in \{1,...,N\}$. All patches are divided into patches of the novel classes foreground and other patches. Their positions are denoted as $P_f \in \{1,...,N\}^J$ and $P_b \in \{1,...,N\}^K$ respectively, where J and K represent the number of foreground and other patches respectively, and $K+J=N$. Then we randomly mask the proportion of the foreground patch of novel classes. The masked positions are denoted $M_f \in \{1,...,N\}^{k\%J}$, where $k\%$ is the masking ratio of foreground patches. The remaining patches x_M^k are shown in Eq. (3).

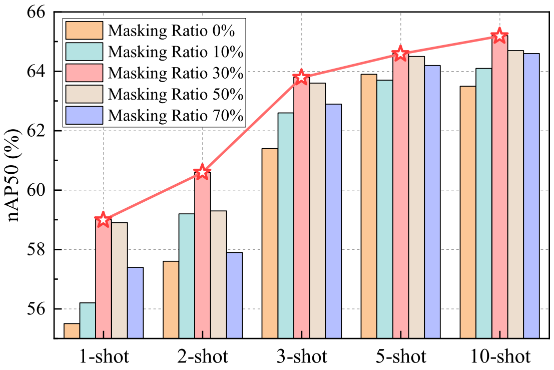

Figure 5. Different masking ratios in OREM on novel split 1 of PASCAL VOC.

$$x_M^k = \{x_i^p : i \in P_b\}_{i=1}^N \cup \{x_i^p : i \notin M_f \cap i \in P_f\}_{i=1}^N \quad (3)$$

$$I_{rec}^k = f_D(f_E(x_M^k)) \quad (4)$$

As illustrated in Eq. (4), x_M^k are sent to the encoder f_E and decoder f_D of MAE to generate the reconstructed image of the novel class I_n^k, which can serve as the enhanced version of the object. Finally, we paste the enhanced object I_{rec}^k back to I_{sal} to generate I_{rec} of d_{rec}.

Figure 5 shows the influence of setting different masking ratios of MAE. And best performance is achieved at the masking ratio of 30%. The higher masking ratio should enhance the diversity of samples. However, the performance diminishes. The observed phenomenon is attributed to the network exhibiting emphasis on low-level image features, resulting in overlooking semantic awareness. To address this issue, we introduce the pre-trained language encoder to constrain the intermediate representation to enhance the semantic awareness of the masked autoencoder. Specifically, it leverages high-level features of CLIP text embedding [35] to guide the semantic features of vision. First, given labels from novel categories of datasets, we use an available CLIP method to represent these labels into rich semantic knowledge as $f_{C_{novel}}$. A linear layer transforms the visible representation f_E generated by MAE's encoder to simulate f_v. The semantic discriminator loss $\mathcal{L}_{SD}$ is defined as follows:

$$\mathcal{L}_{SD} = -\frac{f_{C_{novel}} \cdot linear(f_E)}{\|f_{C_{novel}}\| \cdot \|linear(f_E)\|} + \lambda \|W_v\|_2^2 \quad (5)$$

Where $linear$ is a linear layer. $\cdot$ means the dot product and $\|\cdot\|$ is L2 normalization. The second term is a weight decay term of the MAE's encoder and helps prevent overfitting. Our OREM equipped with $\mathcal{L}_{SD}$ significantly preserves the invariance of high-level semantics in reconstructed images.

Method/Shot	Novel Split 1					Novel Split 2					Novel Split 3					Mean
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FSRW [20] *ICCV 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	28.4
MetaDet [44] *ICCV 19*	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	31.0
TFA w/ cos [45] *ICML 20*	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
MPSR [48] *ECCV 20*	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7	35.8
FSCE [41] *CVPR 21*	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5	46.6
SRR-FSD [58] *CVPR 21*	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	44.8
TIP [26] *CVPR 21*	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9	38.5
FSOD-UP [47] *ICCV 21*	43.8	47.8	50.3	55.4	61.7	31.2	30.5	41.2	42.2	48.3	35.5	39.7	43.9	50.6	53.5	45.0
FADI [2] *NeurIPS 21*	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6	49.2
FCT [14] *CVPR 22*	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	51.6
LVC [22] *CVPR 22*	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	52.3
TENET [56] *ECCV 22*	46.7	-	55.4	62.3	66.9	40.3	-	44.7	49.3	52.1	35.5	-	46.0	54.4	54.6	50.7
MRSN [32] *ECCV 22*	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	47.7
FewX [8] *ECCV 22*	40.1	44.2	51.2	62.0	63.0	33.3	33.1	42.3	46.3	52.3	36.1	43.1	43.5	52.0	56.0	46.6
D&R [28] *AAAI 23*	41.0	51.7	55.7	61.8	65.4	30.7	39.0	42.5	46.6	51.7	37.9	47.1	51.7	56.8	59.5	49.3
ICPE [31] *AAAI 23*	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	50.9	63.1	55.3	60.6	60.1	55.0
VFA [15] *AAAI 23*	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9	56.1
FS-DETR [1] *ICCV 23*	45.0	48.5	51.5	52.7	56.1	37.3	41.3	43.4	46.6	49.0	43.8	47.1	50.6	52.1	56.9	48.1
Du et al. [15] *ICCV 23*	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	55.4
Norm-VAE [51] *CVPR 23*	62.1	64.9	67.8	69.2	67.5	39.9	46.8	54.4	54.2	53.6	58.2	60.3	61.0	64.0	65.5	59.1
DeFRCN [34] *ICCV 21*	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.2	54.9	57.4	51.9
SNIDA-DeFRCN	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**35.2**	**40.8**	**50.2**	**54.6**	**50.0**	**51.6**	**52.4**	**55.9**	**58.5**	**62.6**	**55.1**
MFD [49] *ECCV 22*	63.4	66.3	67.7	69.4	68.1	42.1	46.5	53.4	55.3	53.8	56.1	58.3	59.0	62.2	63.7	59.0
SNIDA-MFD	**64.9**	**67.9**	**69.7**	**71.4**	**70.5**	**42.2**	**47.8**	**54.5**	**56.6**	**54.9**	**58.1**	**61.3**	**60.7**	**63.6**	**66.0**	**60.7**

Table 1. Few-shot detection performance across the 3 splits on the PASCAL VOC benchmark. The best and second-best results are colored red and blue, respectively. Two methods equipped with our idea consistently outperform relevant baselines under all settings and achieve competitive performance compared with recent state-of-the-art methods for 3 novel splits.

4. Experiments

4.1. Few-Shot Object Detection Benchmarks

We followed the previous work [12, 20, 45] to use data splits and novel samples to evaluate the effectiveness of our proposed SNIDA. Two widely used few-shot detection datasets: PASCAL VOC [6] and MS-COCO [29] are adopted to train and evaluate for a fair comparison.

PASCAL VOC have three different data splits, where each split group randomly divides 20 classes into 15 base classes and 5 novel classes. For each novel category, K=1, 2, 3, 5, and 10 shots are available from the combination of VOC2007 and VOC2012 train/val sets for finetuning. We evaluate the performance on the VOC2007 test set with the standard PASCAL VOC metric, Average Precision (IoU=0.5), and report it as nAP50 for the novel categories.

MS-COCO comprises 80 categories, of which 20 are considered novel categories in common with PASCAL VOC, and the remaining 60 belong to the base classes. To finetune the model, we report the outcomes of K=1, 2, 3, 5, 10, and 30 shots for each novel class. We evaluate the performance on 5k images from the validation set. We report the COCO-style AP as the evaluation metric.

4.2. Implementation Details

In this paper, our experiments are conducted on two robust baselines: DeFRCN [34] and MFD [49]. DeFRCN is a simple yet effective finetuning-based framework, that proposes to perform stop-gradient between the RPN and the backbone, and scale-gradient between RCNN and the backbone. MFD is designed to explicitly learn three types of commonalities between base classes and novel classes. Subsequently, these commonalities are extracted during the fine-tuning phase based on a memory bank. We evaluate DeFRCN and MFD performance (mAP) over multiple runs.

Traditional data augmentation is employed in CutMix to prevent overfitting, including random scaling, clipping, and flipping, due to novel classes with only a few samples. We adopt SGD optimization, with the momentum and weight attenuation of 0.9 and 0.0001 respectively. The learning rate during base training is set to 0.02, and the learning rate is set to 0.01 during few-shot finetuning. We pre-train the masked autoencoder on ImageNet-1k [40] for 400 epochs following the setting of [18]. The masking ratio of foreground patches is set to 70%. All experiment was carried out on 4 Nvidia GeForce RTX 3090 GPUs with a batch size of 16 by the open-source library PyTorch (https://pytorch.org/).

Method		Shot Number					
		1	2	3	5	10	30
Meta R-CNN [52]	*ICCV 19*	1.0	1.8	2.8	4.0	6.5	11.1
MPSR [48]	*ECCV 20*	5.1	6.7	7.4	8.7	9.8	14.1
FSDetView [50]	*ECCV 20*	4.5	6.6	7.2	10.7	12.5	14.7
TFA [45]	*ICML 20*	4.4	5.4	6.0	7.7	10.0	13.7
Retentive R-CNN [9]	*CVPR 21*	-	-	-	-	10.5	13.8
CME [27]	*CVPR 21*	-	-	-	-	15.1	16.9
DCNet [19]	*CVPR 21*	-	-	-	-	12.8	18.6
QA-FewDet [12]	*ICCV 21*	4.9	7.6	8.4	9.7	11.6	16.5
FCT [14]	*CVPR 22*	5.6	7.9	11.1	14	17.1	21.4
Meta Faster R-CNN [13]	*AAAI 22*	5.1	7.6	9.8	10.8	12.7	16.6
D&R [28]	*AAAI 23*	8.3	12.7	14.3	16.4	18.7	21.8
FS-DETR [1]	*ICCV 23*	7.0	8.9	10.0	10.9	11.3	-
Du et al. [15]	*ICCV 23*	-	-	-	-	20.3	22.8
Norm-VAE [51]	*CVPR 23*	9.5	13.7	14.3	15.9	18.7	22.5
DeFRCN [34]	*ICCV 21*	9.3	12.9	14.8	16.1	18.5	22.6
SNIDA-DeFRCN		**10.2**	**14.5**	**15.8**	**16.9**	**19.1**	**23.1**
MFD [49]	*ECCV 22*	10.8	13.9	15.0	16.4	19.4	22.7
SNIDA-MFD		**12.0**	**15.4**	**16.4**	**17.8**	**20.7**	**23.8**

Table 2. Experiments on COCO dataset. The best and second-best results are colored red and blue, respectively. Two methods equipped with our idea consistently outperform relevant baselines under all settings. A new state-of-the-art result is also achieved.

4.3. Comparison Results

Results on PASCAL VOC. Table 1 presents the results from 3 novel splits of PASCAL VOC compared with baselines and the existing state-of-the-art methods. Our few-shot data augmentation method surpasses the two baselines in all splits and shots. Our method based on MFD [49] achieves the best performance with 60.7%, which outperforms the baseline with a margin of 1.7 on average. In terms of overall performance, our method illustrates superior performance compared to most existing methods across different splits and shots on average, which demonstrates the robustness and generalization of our method.

Results on MS-COCO. For the MS COCO dataset, we adopt some recent works for comparison, as shown in Table 2. After applying our method, we consistently achieve performance improvement over the two baselines on average respectively. New state-of-the-art results under different shot settings are achieved with other existing methods. Under the setting of 1, 2, 3, 5, 10, and 30 shots, our method achieves 11%, 11%, 9%, 9%, 7%, and 5% performance gain compared with MFD [49] on AP respectively. Especially under the setting of 1-2 shot instances is quite limited, our method is greatly improved with a boost of up to about 10% AP. The results show that our data augmentation method performs better on the more difficult MS COCO dataset.

Comparision among Different Augmentation Methods. Table 3 shows the results of Cutout [5], GridMask [4], CutMix [55], and our method on the Novel Split 1 of PASCAL VOC. We adopt DeFRCN [34] as the baseline. The proposed method is able to consistently outperform other augmentation methods and achieves better performance across different shots. Unlike the above methods, our method obtains semantic awareness through foreground/background decoupling and fusing. The diversity of instances in our method is further enhanced thanks to the adoption of the semantic-guided non-linear transformation.

	Cutout	GridMask	CutMix	Ours
Image				
1-shot	54.7	54.3	55.5	**59.3**
2-shot	57.2	58.0	57.6	**60.8**
3-shot	62.3	62.0	61.4	**64.3**
5-shot	64.0	63.7	63.9	**65.4**
10-shot	62.5	63.2	63.5	**65.6**
Mean	60.1	60.2	60.4	**63.1**

Table 3. The results of Cutout, GridMask, CutMix, and our method on the Novel Split 1 of PASCAL VOC. Our method significantly improves the performance under different shots.

4.4. Ablation Studies

To investigate the proposed method's effectiveness, we conducted three ablation experiments. DeFRCN [34] is adopted as the baseline on the Novel Split 1 of PASCAL VOC. Comprehensive Results on all Novel Splits are provided in the supplementary materials. Table 4 summarizes the results of our method under different settings, and the following is a detailed comparison. Note that we adopt CutMix to increase the number of training samples for novel classes and improve the performance by 1.2 on average.

Impact of Semantic Awareness Enhancement Strategy. To assess the impact of the SAES, we incorporated SAES into the baseline model with CutMix and conducted a comparative analysis. The results presented in Table 4 (line 2 vs line 3) demonstrate that SAES has yielded a considerable improvement, indicating that the decoupling of foreground and background is critical for effectively learning objective semantic discriminative features. By fusing instances into different backgrounds, the adoption of SAES can indeed bring performance improvements.

Impact of Object Reconstruction Enhancement Module. To validate the impact of the OREM, we add and compare it to the baseline with CutMix while keeping other conditions unchanged. The experimental results show that the OREM outperforms the baseline with CutMix and improves the overall performance by 1.4 on average (Table 4, line 2 vs line 4). That is, such a semantic-guided non-linear data enhancement procedure is important for further promoting

CutMix	SAES	OREM	Novel Split 1					Novel Split 2					Novel Split 3					Mean	Speed (s/iter)
			1	2	3	5	10	1	2	3	5	10	1	2	3	5	10		
			52.8	57.7	59.9	62.4	60.0	30.7	38.7	46.9	52.8	47.9	47.3	50.2	52.5	55.2	58.9	51.6	0.48
✓			55.5	57.6	61.4	63.9	63.5	32.0	38.9	47.4	53.6	48.9	48.8	49.9	53.4	56.7	59.8	52.8	0.50
✓	✓		57.3	59.9	63.7	64.1	63.7	33.7	39.3	47.1	54.4	49.6	49.4	51.3	55.2	58.0	61.0	53.9	0.59
✓		✓	58.5	60.1	63.7	64.7	64.5	32.8	39.7	49.2	55.0	49.9	50.0	51.2	55.5	57.9	61.1	54.2	0.57
✓	✓	✓	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**35.2**	**40.8**	**50.2**	**54.6**	**50.0**	**51.6**	**52.4**	**55.9**	**58.5**	**62.6**	**55.1**	0.63

Table 4. Ablation study of different components of our method for FSOD on 3 novel splits of PASCAL VOC. After finetuning with the SAES, our model achieves a significant improvement over the baseline. The performance could be further promoted with OREM adoption.

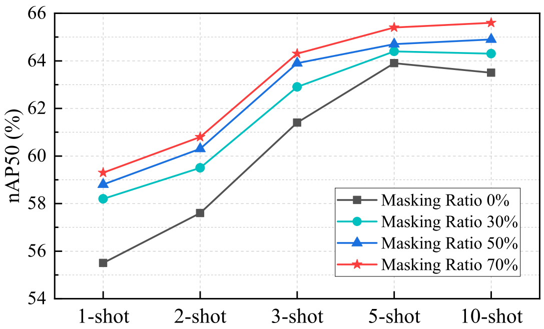

Figure 6. Ablation study of different masking ratios in OREM on novel split 1 of PASCAL VOC. Note the masking ratio here is not for all patches of the entire image, but only for the foreground.

the performance of FSOD since it could better exploit the potential of data by augmenting data diversity.

Moreover, we also carefully analyze the influence of setting different masking ratios $k\%$ in the OREM in Figure 6. It shows that the OREM achieves the best performance at the masking ratio of 70%. Images generated by an undersized masking ratio lack variety, which affects the generalization of the detection model. Compared with Figure 5, the higher masking ratio achieves better performance, which is attributed to the correct semantic-guided high-level semantic supervision imposed by the semantic discrimination loss.

Impact of Different Semantic Supervision. To validate the impact of the different semantic supervision, we conducted experiments within the semantic discriminator of OREM, involving three distinct types of semantic supervision. Firstly, *Random* supervision assigns random and alternative labels to the sample. Secondly, *Misleading* supervision provides consistent but incorrect labels. Thirdly, *Consistency* supervision guides with correct labels. The results indicate that both the first and second supervision resulted in a slight drop in performance compared to the baseline, while the third, involving correct semantic guidance, significantly improved overall performance. This implies that correct high-level supervision aids in preserving the invariance of high-level semantics in reconstructed images. Simultaneously, the introduction of rich non-linear semantic knowledge further enhances the quality and expressive capacity of reconstructed images under high masking ratios.

Semantic Supervision	Shot					Mean
	1	2	3	5	10	
Baseline	57.3	59.9	63.7	64.1	63.7	61.7
Random	55.9	57.8	62.3	62.5	62.4	60.2
Misleading	56.5	58.2	63.1	63.0	63.1	60.8
Consistency	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**63.1**

Table 5. Different semantic supervision in Semantic Discriminator of OREM on novel split 1 of PASCAL VOC.

4.5. Drawback and Discussion

As shown in the last column of Table 4, we also evaluate the speed of finetuning under all five settings. We can see that despite the significant performance improvement, the proposed method is nearly 31% slower than the baseline during the finetuning. However, it is worth noting that finetuning-based FSOD models equipped with our idea will not bring any extra computational burden during the inference but consistently achieve better results.

5. Conclusion

In this paper, we propose an instance-level data augmentation method for FSOD, which decouples the foreground and background to increase their semantic diversity respectively. To decouple the foreground background of few-shot instances, the semantic awareness enhancement strategy is adopted to extract the mask of instances and fuse instances into different backgrounds. The object reconstruction enhancement module is additionally introduced to augment instances in a semantic-guided non-linear manner to further enhance the instance diversity. The proposed method can be readily plugged into existing finetuning-based FSOD methods for further performance improvement. Extensive experiments on two datasets indicate the model equipped with our method significantly outperforms baselines by a large margin. New state-of-the-art results on the PASCAL VOC and MS-COCO under different shot settings are achieved, demonstrating the effectiveness of our idea.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (NSFC) grant 62176100 and 62376011. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

[1] Adrian Bulat, Ricardo Guerrero, Brais Martinez, and Georgios Tzimiropoulos. Fs-detr: Few-shot detection transformer with prompting and without re-training. In *ICCV*, 2023. 6, 7

[2] Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, and Dahua Lin. Few-shot object detection via association and discrimination. In *NeurIPS*, 2021. 6

[3] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *AAAI*, 2018. 3

[4] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 2, 7

[5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 7

[6] Mark Everingham, S. M. Eslami, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 6

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020. 3

[8] Qi Fan, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with model calibration. In *ECCV*, 2022. 6

[9] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *CVPR*, 2021. 1, 7

[10] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[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 *ICCV*, 2021. 6, 7

[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 *AAAI*, 2022. 7

[14] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *CVPR*, 2022. 6, 7

[15] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via variational feature aggregation. *AAAI*, 2023. 6, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 5, 6

[19] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. In *CVPR*, 2021. 7

[20] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 3, 4, 6

[21] 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 *CVPR*, 2019. 3

[22] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few shot object detection method. In *CVPR*, 2022. 2, 3, 6

[23] Geonuk Kim, Hong-Gyu Jung, and Seong-Whan Lee. Spatial reasoning for few-shot object detection. *Pattern Recognition*, 120:108–118, 2021. 2, 3

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 60(6):84–90, 2017. 2

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[26] Aoxue Li and Zhenguo Li. Transformation invariant few-shot object detection. In *CVPR*, 2021. 2, 3, 6

[27] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *CVPR*, 2021. 7

[28] Jiangmeng Li, Yanan Zhang, Wenwen Qiang, Lingyu Si, Chengbo Jiao, Xiaohui Hu, Changwen Zheng, and Fuchun Sun. Disentangle and remerge: interventional knowledge distillation for few-shot object detection from a conditional causal perspective. In *AAAI*, 2023. 6, 7

[29] 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 *ECCV*, 2014. 6

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[31] 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 *AAAI*, 2023. 6

[32] 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 *ECCV*, 2022. 6

[33] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *NeurIPS*, 2019. 4

[34] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *ICCV*, 2021. 1, 3, 4, 6, 7

[35] 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 *ICML*, 2021. 5

[36] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 3

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CVPR*, 2018. 3

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 3

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV (IJCV)*, 115(3): 211–252, 2015. 6

[41] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, 2021. 2, 3, 6

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 2

[43] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, 2021. 4

[44] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *CVPR*, 2019. 6

[45] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020. 1, 3, 4, 6, 7

[46] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *ICCV*, 2019. 3

[47] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *ICCV*, 2021. 6

[48] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *ECCV*, 2020. 2, 3, 6, 7

[49] Shuang Wu, Wenjie Pei, Dianwen Mei, Fanglin Chen, Jiandong Tian, and Guangming Lu. Multi-faceted distillation of base-novel commonality for few-shot object detection. In *ECCV*, 2022. 3, 6, 7

[50] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, 2020. 3, 4, 7

[51] Jingyi Xu, Hieu Le, and Dimitris Samaras. Generating features with increased crop-related diversity for few-shot object detection. In *CVPR*, 2023. 6, 7

[52] 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 *ICCV*, 2019. 7

[53] Yukuan Yang, Fangyun Wei, Miaojing Shi, and Guoqi Li. Restoring negative information in few-shot object detection. In *NeurIPS*, 2020. 3

[54] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: tackling object confusion for few-shot detection. In *AAAI*, 2020. 3

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 4, 7

[56] Shan Zhang, Naila Murray, Lei Wang, and Piotr Koniusz. Time-reversed diffusion tensor transformer: A new tenet of few-shot object detection. In *ECCV*, 2022. 6

[57] Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. In *CVPR*, 2021. 3

[58] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-stable few-shot object detection. In *CVPR*, 2021. 6