Generating Features with Increased Crop-related Diversity for Few-Shot Object Detection

Jingyi Xu
Stony Brook University
jingyixu@cs.stonybrook.edu

Hieu Le
EPFL
minh.le@epfl.ch

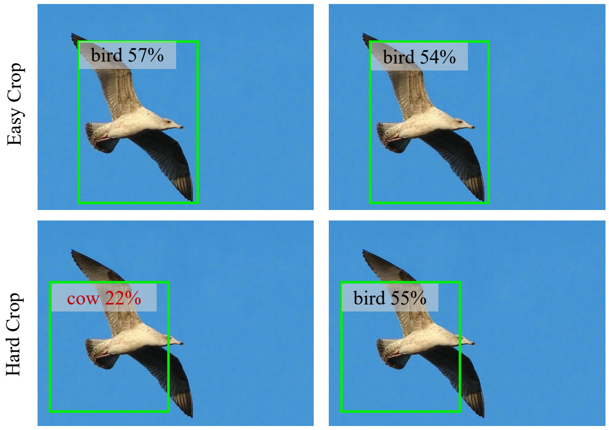
Dimitris Samaras
Stony Brook University
samaras@cs.stonybrook.edu

Abstract

Two-stage object detectors generate object proposals and classify them to detect objects in images. These proposals often do not contain the objects perfectly but overlap with them in many possible ways, exhibiting great variability in the difficulty levels of the proposals. Training a robust classifier against this crop-related variability requires abundant training data, which is not available in few-shot settings. To mitigate this issue, we propose a novel variational autoencoder (VAE) based data generation model, which is capable of generating data with increased crop-related diversity. The main idea is to transform the latent space such latent codes with different norms represent different crop-related variations. This allows us to generate features with increased crop-related diversity in difficulty levels by simply varying the latent norm. In particular, each latent code is rescaled such that its norm linearly correlates with the IoU score of the input crop w.r.t. the ground-truth box. Here the IoU score is a proxy that represents the difficulty level of the crop. We train this VAE model on base classes conditioned on the semantic code of each class and then use the trained model to generate features for novel classes. In our experiments our generated features consistently improve state-of-the-art few-shot object detection methods on the PASCAL VOC and MS COCO datasets.

1. Introduction

Object detection plays a vital role in many computer vision systems. However, training a robust object detector often requires a large amount of training data with accurate bounding box annotations. Thus, there has been increasing attention on few-shot object detection (FSOD), which learns to detect novel object categories from just a few annotated training samples. It is particularly useful for problems where annotated data can be hard and costly to obtain such as rare medical conditions [31, 41], rare animal species [20, 44], satellite images [2, 19], or failure cases in autonomous driving systems [27, 28, 36].



(a) DeFRCN [33]   (b) Ours

Figure 1. **Robustness to different object crops of the same object instance**. (a) The classifier head of the state-of-the-art FSOD method [33] classifies correctly a simple crop of the bird but misclassifies a hard crop where some parts are missing. (b) Our method can handle this case since it is trained with additional generated features with increased crop-related diversity. We show the class with the highest confidence score.

For the most part, state-of-the-art FSOD methods are built on top of a two-stage framework [35], which includes a region proposal network that generates multiple image crops from the input image and a classifier that labels these proposals. While the region proposal network generalizes well to novel classes, the classifier is more error-prone due to the lack of training data diversity [40]. To mitigate this issue, a natural approach is to generate additional features for novel classes [12, 55, 57]. For example, Zhang *et al.* [55] propose a feature hallucination network to use the variation from base classes to diversify training data for novel classes. For zero-shot detection (ZSD), Zhu *et al.* [57] propose to synthesize visual features for unseen objects based on a conditional variational auto-encoder. Although much progress has been made, the lack of data diversity is still a challenging issue for FSOD methods.

Here we discuss a specific type of data diversity that greatly affects the accuracy of FSOD algorithms. Specifically, given a test image, the classifier needs to accurately classify multiple object proposals[1] that overlap the object instance in various ways. The features of these image crops exhibit great variability induced by different object scales, object parts included in the crops, object positions within the crops, and backgrounds. We observe a typical scenario where the state-of-the-art FSOD method, DeFRCN [33], only classifies correctly a few among many proposals overlapping an object instance of a few-shot class. In fact, different ways of cropping an object can result in features with various difficulty levels. An example is shown in Figure 1a where the image crop shown in the top row is classified correctly while another crop shown in the bottom row confuses the classifier due to some missing object parts. In general, the performance of the method on those hard cases is significantly worse than on easy cases (see section 5.4). However, building a classifier robust against crop-related variation is challenging since there are only a few images per few-shot class.

In this paper, we propose a novel data generation method to mitigate this issue. Our goal is to generate features with diverse crop-related variations for the few-shot classes and use them as additional training data to train the classifier. Specifically, we aim to obtain a diverse set of features whose difficulty levels vary from easy to hard *w.r.t.* how the object is cropped.[2] To achieve this goal, we design our generative model such that it allows us to control the difficulty levels of the generated samples. Given a model that generates features from a latent space, our main idea is to enforce that the magnitude of the latent code linearly correlates with the difficulty level of the generated feature, *i.e.*, the latent code of a harder feature is placed further away from the origin and vice versa. In this way, we can control the difficulty level by simply changing the norm of the corresponding latent code.

In particular, our data generation model is based on a conditional variational autoencoder (VAE) architecture. The VAE consists of an encoder that maps the input to a latent representation and a decoder that reconstructs the input from this latent code. In our case, inputs to the VAE are object proposal features, extracted from a pre-trained object detector. The goal is to associate the norm (magnitude) of the latent code with the difficulty level of the object proposal. To do so, we rescale the latent code such that its norm linearly correlates with the Intersection-Over-Union (IoU) score of the input object proposal *w.r.t.* the ground-truth object box. This IoU score is a proxy that partially indicates the difficulty level: A high IoU score indicates that the object proposal significantly overlaps with the object instance while a low IoU score indicates a harder case where a part of the object can be missing. With this rescaling step, we can bias the decoder to generate harder samples by increasing the latent code magnitude and vice versa. In this paper, we use latent codes with different norms varying from small to large to obtain a diverse set of features which can then serve as additional training data for the few-shot classifier.

To apply our model to FSOD, we first train our VAE model using abundant data from the base classes. The VAE is conditioned on the semantic code of the input instance category. After the VAE model is trained, we use the semantic embedding of the few-shot class as the conditional code to synthesize new features for the corresponding class. In our experiments, we use our generated samples to fine-tune the baseline few-shot object detector - DeFRCN [33]. Surprisingly, a vanilla conditional VAE model trained with only ground-truth box features brings a 3.7% nAP50 improvement over the DeFRCN baseline in the 1-shot setting of the PASCAL VOC dataset [4]. Note that we are the first FSOD method using VAE-generated features to support the training of the classifier. Our proposed Norm-VAE can further improve this new state-of-the-art by another 2.1%, *i.e.*, from 60% to 62.1%. In general, the generated features from Norm-VAE consistently improve the state-of-the-art few-shot object detector [33] for both PASCAL VOC and MS COCO [24] datasets.

Our main contributions can be summarized as follows:

- We show that lack of crop-related diversity in training data of novel classes is a crucial problem for FSOD.
- We propose Norm-VAE, a novel VAE architecture that can effectively increase crop-related diversity in difficulty levels into the generated samples to support the training of FSOD classifiers.
- Our experiments show that the object detectors trained with our additional features achieve state-of-the-art FSOD in both PASCAL VOC and MS COCO datasets.

2. Related Work

Few-shot Object Detection Few-shot object detection aims to detect novel classes from limited annotated examples of previously unseen classes. A number of prior methods [5, 7, 8, 10, 11, 17, 17, 21, 23, 25, 26, 32, 40, 45–47, 56] have been proposed to address this challenging task. One line of work focuses on the **meta-learning** paradigm, which has been widely explored in few-shot classification [6, 16, 37, 43, 50, 52–54]. Meta-learning based approaches introduce a meta-learner to acquire meta-knowledge that can be then transferred to novel classes. [16] propose a meta feature learner and a reweighting module to fully exploit generalizable features from base classes and quickly adapt the prediction network to predict novel classes. [43] pro-

[1] Note that an RPN typically outputs 1000 object proposals per image.

[2] In this paper, the difficulty level is strictly related to how the object is cropped.

pose specialized meta-strategies to disentangle the learning of category-agnostic and category-specific components in a CNN based detection model. Another line of work adopts a **two-stage fine-tuning** strategy and has shown great potential recently [3,33,40,42,48]. [42] propose to fine-tune only box classifier and box regressor with novel data while freezing the other paramters of the model. This simple stragetegy outperforms previous meta-learners. FSCE [40] leverages a contrastive proposal encoding loss to promote instance level intra-class compactness and inter-class variance. Orthogonal to existing work, we propose to generate new samples for FSOD. Another **data generation based** method for FSOD is Halluc [55]. However, their method learns to transfer the shared within-class variation from base classes while we focus on the crop-related variance.

Feature Generation Feature generation has been widely used in low-shot learning tasks. The common goal is to generate reliable and diverse additional data. For example, in image classification, [51] propose to generate representative samples using a VAE model conditioned on the semantic embedding of each class. The generated samples are then used together with the original samples to construct class prototypes for few-shot learning. In spirit, their conditional-VAE system is similar to ours. [49] propose to combine a VAE and a Generative Adversarial Network (GAN) by sharing the decoder of VAE and generator of GAN to synthesize features for zero-shot learning. In the context of object detection, [55] propose to transfer the shared modes of within-class variation from base classes to novel classes to hallucinate new samples. [56] propose to synthesize visual features for unseen objects from semantic information and augment existing training algorithms to incorporate unseen object detection. Recently, [15] propose to synthesize samples which are both intra-class diverse and inter-class separable to support the training of zero-shot object detector. However, these methods do not take into consideration the variation induced by different crops of the same object, which is the main focus of our proposed method.

Variational Autoencoder Different VAE variants have been proposed to generate diverse data [9, 14, 18, 38]. β-VAE [14] imposes a heavy penalty on the KL divergence term to enhance the disentanglement of the latent dimensions. By traversing the values of latent variables, β-VAE can generate data with disentangled variations. ControlVAE [38] improves upon β-VAE by introducing a controller to automatically tune the hyperparameter added in the VAE objective. However, disentangled representation learning can not capture the desired properties without supervision. Some VAE methods allow explicitly controllable feature generation including CSVAE [18] and PCVAE [9]. CSVAE [18] learns latent dimensions associated with binary properties. The learned latent subspace can easily be inspected and independently manipulated. PCVAE [9] uses a Bayesian model to inductively bias the latent representation. Thus, moving along the learned latent dimensions can control specific properties of the generated data. Both CSVAE and PCVAE use additional latent variables and enforce additional constrains to control properties. In contrast, our Norm-VAE directly encodes a variational factor into the norm of the latent code. Experiments show that our strategy outperforms other VAE architectures, while being simpler and without any additional training components.

3. Method

In this section, we first review the problem setting of few-shot object detection and the conventional two-stage fine-tuning framework. Then we introduce our method that tackles few-shot object detection via generating features with increased crop-related diversity.

3.1. Preliminaries

In few-shot object detection, the training set is divided into a base set D^B with abundant annotated instances of classes C^B, and a novel set D^N with few-shot data of classes C^N, where C^B and C^N are non-overlapping. For a sample $(x, y) \in D^B \cup D^N$, x is the input image and $y = \{(c_i, b_i), i = 1, ..., n\}$ denotes the categories $c \in C^B \cup C^N$ and bounding box coordinates b of the n object instances in the image x. The number of objects for each class in C^N is K for K-shot detection. We aim to obtain a few-shot detection model with the ability to detect objects in the test set with classes in $C^B \cup C^N$.

Recently, two-stage fine-tuning methods have shown great potential in improving few-shot detection. In these two-stage detection frameworks, a Region Proposal Network (RPN) takes the output feature maps from a backbone feature extractor as inputs and generates region proposals. A Region-of-Interest (RoI) head feature extractor first pools the region proposals to a fixed size and then encodes them as vector embeddings, known as the RoI features. A classifier is trained on top of the RoI features to classify the categories of the region proposals.

The fine-tuning often follows a simple two-stage training pipeline, *i.e.*, the data-abundant base training stage and the novel fine-tuning stage. In the base training stage, the model collects transferable knowledge across a large base set with sufficient annotated data. Then in the fine-tuning stage, it performs quick adaptation on the novel classes with limited data. Our method aims to generate features with diverse crop-related variations to enrich the training data for the classifier head during the fine-tuning stage. In our experiments, we show that our generated features significantly improve the performance of DeFRCN [33].

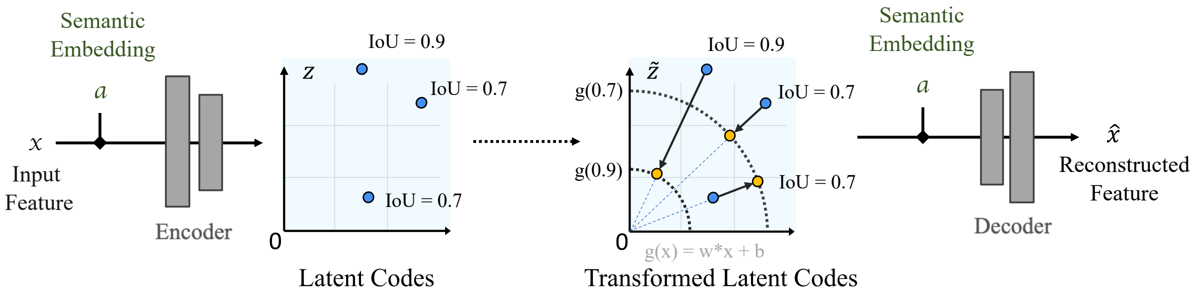


Figure 2. **Norm-VAE for modelling crop-related variations.** The original latent code z is rescaled to $\hat{z}$ such that the norm of $\hat{z}$ linearly correlates with the IoU score of the input crop (*w.r.t.* the ground truth box). The original latent codes are colored in **blue** while the rescaled ones are colored in **yellow**. The norm of the new latent code is the output of a simple linear function $g(\cdot)$ taking the IoU score as the single input. As can be seen, the two points whose IoU = 0.7 are both rescaled to norm $g(0.7)$ while another point whose IoU = 0.9 is mapped to norm $g(0.9)$. As a result, different latent norms represent different crop-related variations, enabling diverse feature generation.

3.2. Overall Pipeline

Figure 2 summarizes the main idea of our proposed VAE model. For each input object crop, we first use a pre-trained object detector to obtain its RoI feature. The encoder takes as input the RoI feature and the semantic embedding of the input class to output a latent code z. We then transform z such that its norm linearly correlates with the IoU score of the input object crop *w.r.t.* the ground-truth box. The new norm is the output of a simple linear function $g(\cdot)$ taking the IoU score as the single input. The decoder takes as input the new latent code and the class semantic embedding to output the reconstructed feature. Once the VAE is trained, we use the semantic embedding of the few-shot class as the conditional code to synthesize new features for the class. To ensure the diversity *w.r.t.* object crop in generated samples, we vary the norm of the latent code when generating features. The generated features are then used together with the few-shot samples to fine-tune the object detector.

3.2.1 Norm-VAE for Feature Generation

We develop our feature generator based on a conditional VAE architecture [39]. Given an input object crop, we first obtain its Region-of-Interest (RoI) feature f via a pre-trained object detector. The RoI feature f is the input for the VAE. The VAE is composed of an Encoder $E(f, a)$, which maps a visual feature f to a latent code z, and a decoder $G(z, a)$ which reconstructs the feature f from z. Both E and G are conditioned on the class semantic embedding a. We obtain this class semantic embedding a by inputting the class name into a semantic model [30, 34]. It contains class-specific information and serves as a controller to determine the categories of the generated samples. Conditioning on these semantic embeddings allows reliably generating features for the novel classes based on the learned information from the base classes [51]. Here we assume that the class names of both base and novel classes are available and we can obtain the semantic embedding of all classes.

We first start from a vanilla conditional VAE model. The loss function for training this VAE for a feature f_i of class j can be defined as:

$$L_V(f_i) = \text{KL}\left(q(z_i|f_i, a^j)||p(z|a^j)\right) - \text{E}_{q(z_i|f_i,a^j)}[\log p(f_i|z_i, a^j)], \tag{1}$$

where a^j is the semantic embedding of class j. The first term is the Kullback-Leibler divergence between the VAE posterior $q(z|f, a)$ and a prior distribution $p(z|a)$. The second term is the decoder's reconstruction error. $q(z|f, a)$ is modeled as $E(f, a)$ and $p(f|z, a)$ is equal to $G(z, a)$. The prior distribution is assumed to be $\mathcal{N}(0, I)$ for all classes.

The goal is to control the crop-related variation in a generated sample. Thus, we establish a direct correspondence between the latent norm and the crop-related variation. To accomplish this, we transform the latent code such that its norm correlates with the IoU score of the input crop. Given an input RoI feature f_i of a region with an IoU score s_i, we first input this RoI feature to the encoder to obtain its latent code z_i. We then transform z_i to $\tilde{z}_i$ such that the norm of $\tilde{z}_i$ correlates to s_i. The new latent code $\tilde{z}_i$ is the output of the transformation function $\mathcal{T}(\cdot, \cdot)$:

$$\tilde{z}_i = \mathcal{T}(z_i, s_i) = \frac{z_i}{\|z_i\|} * g(s_i), \tag{2}$$

where $\|z_i\|$ is the L_2 norm of z_i, s_i is the IoU score of the input proposal *w.r.t.* its ground-truth object box, and $g(\cdot)$ is a simple pre-defined linear function that maps an IoU score to a norm value. With this new transformation step, the loss function of the VAE from equation 1 for an input feature f_i from class j with an IoU score s_i thus can be rewritten as:

$$L_V(f_i, s_i) = \text{KL}\left(q(z_i|f_i, a^j)||p(z|a^j)\right) - \text{E}_{q(z_i|f_i,a^j)}\left[\log p(f_i|\mathcal{T}(z_i, s_i), a^j)\right]. \tag{3}$$

3.2.2 Generating Diverse Data for Improving Few-shot Object Detection

After the VAE is trained on the base set, we generate a set of features with the trained decoder. Given a class y with a semantic vector a^y and a noise vector z, we generate a set of augmented features $\mathbb{G}^y$:

$$\mathbb{G}^y = \{\hat{f} | \hat{f} = G(\frac{z}{||z||} * \beta, a^y)\}, \quad (4)$$

where we vary β to obtain generated features with more crop-related variations. The value range of β is chosen based on the mapping function $g(\cdot)$. The augmented features are used together with the few-shot samples to fine-tune the object detector. We fine-tune the whole system using an additional classification loss computed on the generated features together with the original losses computed on real images. This is much simpler than the previous method of [55] where they fine-tune their system via an EM-like (expectation-maximization) manner.

4. Experiments

4.1. Datasets and Evaluation Protocols

We conduct experiments on both PASCAL VOC (07 + 12) [4] and MS COCO datasets [24]. For fair comparison, we follow the data split construction and evaluation protocol used in previous works [16]. The PASCAL VOC dataset contains 20 categories. We use the same 3 base/novel splits with TFA [42] and refer them as Novel Split 1,2, 3. Each split contains 15 base classes and 5 novel classes. Each novel class has K annotated instances, where $K = 1, 2, 3, 5, 10$. We report AP50 of the novel categories (nAP50) on VOC07 test set. For MS COCO, the 60 categories disjoint with PASCAL VOC are used as base classes while the remaining 20 classes are used as novel classes. We evaluate our method on shot 1,2,3,5,10,30 and COCO-style AP of the novel classes is adopted as the evaluation metrics.

4.2. Implementation Details

Feature generation methods like ours in theory can be built on top of many few-shot object detectors. In our experiments, we use the pre-trained Faster-RCNN [35] with ResNet-101 [13] following previous work DeFRCN [33]. The dimension of the extracted RoI feature is 2048. For our feature generation model, the encoder consists of three fully-connected (FC) layers and the decoder consists of two FC layers, both with 4096 hidden units. LeakyReLU and ReLU are the non-linear activation functions in the hidden and output layers, respectively. The dimensions of the latent space and the semantic vector are both set to be 512. Our semantic embeddings are extracted from a pre-trained CLIP [34] model in all main experiments. An additional experiment using Word2Vec [29] embeddings is reported in Section 5.2. After the VAE is trained on the base set with various augmented object boxes , we use the trained decoder to generate $k = 30$ features per class and incorporate them into the fine-tuning stage of the DeFRCN model. We set the function $g(\cdot)$ in Equation 2 to a simple linear function $g(x) = w * x + b$ which maps an input IoU score x to the norm of the new latent code. Note that x is in range $[0.5, 1]$ and the norm of the latent code of our VAE before the rescaling typically centers around $\sqrt{512}$ (512 is the dimension of the latent code). We empirically choose $g(\cdot)$ such that the new norm ranges from $\sqrt{512}$ to $5 * \sqrt{512}$. We provide further analyses on the choice of $g(\cdot)$ in the supplementary material. For each feature generation iteration, we gradually increase the value of the controlling parameter β in Equation 4 with an interval of 0.75.

4.3. Few-shot Detection Results

We use the generated features from our VAE model together with the few-shot samples to fine-tune DeFRCN. We report the performance of two models: “Vanilla-VAE” denotes the performance of the model trained with generated features from a vanilla VAE trained on the base set of ground-truth bounding boxes and “Norm-VAE” denotes the performance of the model trained with features generated from our proposed Norm-VAE model.

PASCAL VOC Table 1 shows our results for all three random novel splits from PASCAL VOC. Simply using a VAE model trained with the original data outperforms the state-of-the-art method DeFRCN in all shot and split on PASCAL VOC benchmark. In particular, vanilla-VAE improves DeFRCN by 3.7% for 1-shot and 4.3% for 3-shot on Novel Split 1. Using additional data from our proposed Norm-VAE model consistently improves the results across all settings. We provide qualitative examples in the supplementary material.

MS COCO Table 2 shows the FSOD results on MS COCO dataset. Our generated features bring significant improvements in most cases, especially in low-shot settings (K $\leq$ 10). For example, Norm-VAE brings a 2.9% and a 2.0% nAP improvement over DeFRCN in 1-shot and 2-shot settings, respectively. Pseudo-Labeling is better than our method in higher shot settings. However, they apply mosaic data augmentation [1] during fine-tuning.

5. Analyses

5.1. Effectiveness of Norm-VAE

We compare the performance of Norm-VAE with a baseline vanilla VAE model that is trained with the same set of augmented data. As shown in Table 4, using the vanilla VAE with more training data does not bring performance improvement compared to the VAE model trained with the

Method	Novel Split 1					Novel Split 2					Novel Split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
TFA w/ fc [42]	36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
TFA w/ cos [42]	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [48]	41.7	-	51.4	55.2	61.8	24.4	-	39.2	35.1	39.9	47.8	-	42.3	48.0	49.7
FsDetView [50]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
FSCE [40]	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
CME [22]	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
SRR-FSD [56]	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
Halluc. [55]	45.1	44.0	44.7	55.0	55.9	23.2	27.5	35.1	34.9	39.0	30.5	35.1	41.4	49.0	49.3
FSOD-MC [5]	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
FADI [3]	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	48.3	51.5
CoCo-RCNN [25]	43.9	44.5	53.1	64.6	65.5	29.4	31.3	43.8	44.3	51.8	39.1	43.9	47.2	54.7	60.3
MRSN [26]	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
FCT [11]	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
Pseudo-Labelling [17]	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
DeFRCN [33]	56.3	60.3	62.0	67.0	66.1	35.7	45.2	51.5	54.1	53.3	54.5	55.6	56.6	60.8	62.7
Vanila-VAE (Ours)	60.0	63.3	66.3	68.3	67.1	39.3	46.2	52.7	53.5	53.4	56.0	58.8	57.1	62.6	63.6
Norm-VAE (Ours)	**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**

Table 1. **Few-shot object detection performance (nAP50) on PASCAL VOC dataset**. We evaluate the performance on three different splits. Our method consistently improves upon the baseline for all three splits across all shots. Best performance in bold.

Method	nAP						nAP75					
	1	2	3	5	10	30	1	2	3	5	10	30
TFA w/ fc [42]	2.9	4.3	6.7	8.4	10.0	13.4	2.8	4.1	6.6	8.4	9.2	13.2
TFA w/ cos [42]	3.4	4.6	6.6	8.3	10.0	13.7	3.8	4.8	6.5	8.0	9.3	13.2
MPSR [48]	2.3	3.5	5.2	6.7	9.8	14.1	2.3	3.4	5.1	6.4	9.7	14.2
FADI [3]	5.7	7.0	8.6	10.1	12.2	16.1	6.0	7.0	8.3	9.7	11.9	15.8
FCT [11]	-	7.9	-	-	17.1	21.4	-	7.9	-	-	17.0	22.1
Pseudo-Labelling [17] †	-	-	-	-	17.8	**24.5**	-	-	-	-	**17.8**	**25.0**
DeFRCN [33]	6.6	11.7	13.3	15.6	18.7	22.4	7.0	12.2	13.6	15.1	17.6	22.2
Vanilla-VAE (ours)	8.8	13.0	14.1	**15.9**	**18.7**	22.5	7.9	12.5	13.4	15.1	17.6	22.2
Norm-VAE (ours)	**9.5**	**13.7**	**14.3**	**15.9**	**18.7**	22.5	**8.8**	**13.7**	**14.2**	**15.3**	**17.8**	22.4

Table 2. **Few-shot detection performance for the novel classes on MS COCO dataset**. Our approach outperforms baseline methods in most cases, especially in low-shot settings ($K < 10$). † applies mosaic data augmentation introduced in [1] during fine-tuning. Best performance in bold.

base set. This suggests that training with more diverse data does not guarantee diversity in generated samples *w.r.t.* a specific property. Our method, by contrast, improves the baseline model by $1.3\% \sim 1.9\%$, which demonstrates the effectiveness of our proposed Norm-VAE.

5.2. Performance Using Different Semantic Embeddings

We use CLIP [34] features in our main experiments. In Table 3, we compare this model with another model trained with Word2Vec [29] on PASCAL VOC dataset. Note that CLIP model is trained with 400M pairs (image and its text title) collected from the web while Word2Vec is trained with only text data. Our Norm-VAE trained with Word2Vec embedding achieves similar performance to the model trained with CLIP embedding. In both cases, the model outperform the state-of-the-art FSOD method in all settings.

5.3. Robustness against Inaccurate Localization

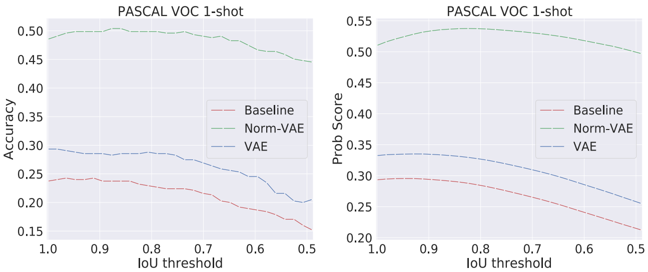
In this section, we conduct experiments to show that our object detector trained with features with diverse crop-related variation is more robust against inaccurate bounding box localization. Specifically, we randomly select 1000 testing instances from PASCAL VOC test set and create 30 augmented boxes for each ground-truth box. Each augmented box is created by enlarging the ground-truth boxes by $x\%$ for each dimension where x ranges from 0 to 30. The result is summarized in Figure 3 where "Baseline" denotes the performance of DeFRCN [33], "VAE" is the performance of the model trained with features generated from a vanilla VAE, and "Norm-VAE" is the model trained with generated features from our proposed model.

Method	Semantic Embedding	Novel Split 1			Novel Split 2			Novel Split 3		
		1-shot	2-shot	3-shot	1-shot	2-shot	3-shot	1-shot	2-shot	3-shot
DeFRCN [33]	-	56.3	60.3	62.0	35.7	45.2	51.5	54.5	55.6	56.6
Vanilla VAE	Word2Vec	60.4	62.9	**66.7**	38.7	45.2	52.9	55.6	58.7	57.9
Norm-VAE		**61.6**	**63.4**	66.3	**40.7**	**46.4**	**53.3**	**56.8**	**59.0**	**60.2**
Vanilla VAE	CLIP	60.0	63.3	66.3	39.3	46.2	52.7	56.0	58.8	57.1
Norm-VAE		**62.1**	**64.9**	**67.8**	**39.9**	**46.8**	**54.4**	**58.2**	**60.3**	**61.0**

Table 3. **FSOD Performance of VAE models trained with different class semantic embeddings**. CLIP [34] is trained with 400M pairs (image and its text title) collected from the web while Word2Vec [29] is trained with only text data.

	Data	1-shot	2-shot	3-shot
DeFRCN [33]	-	56.3	60.3	62.0
VAE	Orginal	60.0	63.3	66.3
VAE	Augmented	60.1	62.7	66.4
Norm-VAE	Augmented	**62.1**	**64.9**	**67.8**

Table 4. **Performance comparisons between vanilla VAE and Norm-VAE on PASCAL VOC dataset**. Training a the vanilla VAE with the augmented data does not bring performance improvement. One possible reason is that the generated samples are not guaranteed to be diverse even with sufficient data.



(a) Accuracy (b) Probability score

Figure 3. **Classification accuracy and probability score of the object detector on the augmented box**. We compare between the baseline DeFRCN [33], the model trained with features from vanilla VAE and our proposed Norm-VAE. By generating features with diverse crop-related variations, we increase the object detector's robustness against inaccurate object box localization.

Method	1-shot	2-shot	3-shot
DeFRCN [33]	16.6	13.3	15.2
Ours (↑ Improvement)	18.8 (↑2.2)	16.4 (↑ 3.1)	19.2 (↑4.0)

Table 5. **AP50∼75 of our method and DeFRCN on PASCAL VOC dataset**. AP 50∼75 refers to the average precision computed on the proposals with the IoU thresholds between 50% and 75% and discard the proposals with IoU scores larger than 0.75, i.e., only "*hard*" cases.

Figure 3 (a) shows the classification accuracy of the object detector on the augmented box as the IoU score between the augmented bounding box and the ground-truth box decreases. For both the baseline method DeFRCN and the model trained with features from a vanilla VAE, the accuracy drops by $\sim$ 10% as the IoU score decreases from 1.0 to 0.5. These results suggest that these models perform much better for boxes that have higher IoU score *w.r.t.* the ground-truth boxes. Our proposed method has higher robustness to these inaccurate boxes: the accuracy of the model trained with features from Norm-VAE only drops by $\sim$ 5% when IoU score decreases from 1 to 0.5.

Figure 3 (b) plots the average probability score of the classifier on the ground-truth category as the IoU score decreases. Similarly, the probability score of both baseline DeFRCN and the model trained with features from a vanilla VAE drops around 0.08 as the IoU score decreases from 1.0 to 0.5. The model trained with features from Norm-VAE, in comparison, has more stable probability score as the IoU threshold decreases.

5.4. Performance on Hard Cases

In Table 5, we show AP 50∼75 of our method on PASCAL VOC dataset (Novel Split 1) in comparison with the state-of-the-art method DeFRCN. Here AP 50∼75 refers to the average precision computed on the proposals with the IoU thresholds between 50% and 75% and discard the proposals with IoU scores (*w.r.t.* the ground-truth box) larger than 0.75. Thus, AP 50∼75 implies the performance of the model in "*hard*" cases where the proposals do not significantly overlap the ground-truth object boxes. In this extreme test, the performance of both models are worse than their AP50 counterparts (Table 1), showing that FSOD methods are generally not robust to those hard cases. Our method mitigates this issue, outperforming DeFRCN by substantial margins. However, the performance is still far from perfect. Addressing these challenging cases is a fruitful venue for future FSOD work.

Features	1-shot		2-shot		3-shot		5-shot	
	nAP50	nAP75	nAP50	nAP75	nAP50	nAP75	nAP50	nAP75
Low-IoU (Hard cases)	**60.9**	30.5	**63.7**	40.6	**66.6**	40.7	**68.9**	41.2
High-IoU (Easy cases)	60.2	**31.6**	63.2	**41.0**	66.3	**41.5**	68.3	**42.1**

Table 6. **Comparison between models trained with different groups of generated features**. The model trained with "Low-IoU" (hard cases) features has better nAP50 scores while the "High-IoU" (easy cases) model has better nAP75 scores. Features corresponding to different difficulty levels improve the performance differently in terms of nAP50 and nAP75.

5.5. Performance with Different Subsets of Generated Features

In this section, we conduct experiments to show that different groups of generated features affect the performance of the object detector differently. Similar to Section 4.2, we generate 30 new features per few-shot class with various latent norms. However, instead of using all norms, we only use large norms (top 30% highest values) to generate the first group of features and only small norms (top 30% lowest values) to generate the second group of features. During training, larger norms correlate to input crops with smaller IoU scores *w.r.t.* the ground-truth boxes and vice versa. Thus, we denote these two groups as "Low-IoU" and "High-IoU" correspondingly. We train two models using these two sets of features and compare their performance in Table 6. As can be seen, the model trained with "Low-IoU" features has higher AP50 while the "High-IoU" model has higher AP75 score. This suggests that different groups of features affect the performance of the classifier differently. The "Low-IoU" features tend to increase the model's robustness to hard-cases while the "High-IoU" features can improve the performance for easier cases. Note that the performance of both of these models is not as good as the model trained with diverse variations and interestingly, very similar to the performance of the vanilla VAE model (Table 1).

5.6. Comparisons with Other VAE architectures

Our proposed Norm-VAE can increase diversity *w.r.t.* image crops in generated samples. Here, we compare the performance of our proposed Norm-VAE with other VAE architectures, including β-VAE [14] and CSVAE [18]. We train all models on image features of augmented object crops on PASCAL VOC dataset using the same backbone feature extractor. For β-VAE, we generate additional features by traversing a randomly selected dimension of the latent code. For CSVAE, we manipulate the learned latent subspace to enforce variations in the generated samples. We use generated features from each method to finetune DeFRCN. The results are summarized in Table 7. In all cases, the generated features greatly benefit the baseline DeFRCN. This shows that lacking crop-related variation is a critical issue for FSOD, and augmenting features with increased crop-related diversity can effectively alleviate the problem. Our proposed Norm-VAE outperforms both β-VAE and CSVAE in all settings. Note that CSVAE requires additional encoders to learn a pre-defined subspace correlated with the property, while our Norm-VAE directly encode this into the latent norm without any additional constraints.

	1-shot	2-shot	3-shot
DeFRCN [33]	56.3	60.3	62.0
β-VAE [14]	61.3	64.0	67.3
CSVAE [18]	61.6	64.1	67.4
Norm-VAE	**62.1**	**64.9**	**67.8**

Table 7. **Comparison between Norm-VAE and other VAE variants.** Norm-VAE outperforms β-VAE and CSVAE on PASCAL VOC dataset under all settings. Best performance in bold.

6. Conclusion and Future Works

We tackle the lack of crop-related variability in the training data of FSOD, which makes the model not robust to different object proposals of the same object instance. To this end, we propose a novel VAE model that can generate features with increased crop-related diversity. Experiments show that such increased diversity in the generated samples significantly improves the current state-of-the-art FSOD performance for both PASCAL VOC and MS COCO datasets. Our proposed VAE model is simple, easy to implement, and allows modifying the difficulty levels of the generated samples. In general, generative models whose outputs can be manipulated according to different properties, are crucial to various frameworks and applications. In future work, we plan to address the following limitations of our work: 1) We bias the decoder to increase the diversity in generated samples instead of explicitly enforcing it. 2) Our proposed method is designed to generate visual features of object boxes for FSOD. Generating images might be required in other applications. Another direction to extend our work is to represent other variational factors in the embedding space to effectively diversify generated data.

Acknowledgements. This research was partially supported by NSF grants IIS-2123920 and IIS-2212046 and the NASA Biodiversity program (Award 80NSSC21K1027).

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020. 5, 6

[2] Alex Borowicz, Hieu Le, Grant Humphries, G. Nehls, Caroline Höschle, V. Kosarev, and H. Lynch. Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. *PLoS ONE*, 14, 2019. 1

[3] 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. 3, 6

[4] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88, 2009. 2, 5

[5] Qi Fan, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with model calibration. In *ECCV*, 2022. 2, 6

[6] Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, pages 4012–4021, 2020. 2

[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*, June 2020. 2

[8] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *CVPR*, June 2021. 2

[9] Xiaojie Guo, Yuanqi Du, and Liang Zhao. Property controllable variational autoencoder via and invertible mutual dependence. In *ICLR*, 2021. 3

[10] 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*, October 2021. 2

[11] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *CVPR*, 2022. 2, 6

[12] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Hameed Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 5

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3, 8

[15] Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust region feature synthesizer for zero-shot object detection. In *CVPR*, 2022. 3

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. *ICCV*, 2019. 2, 5

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

[18] Jack Klys, Jake Snell, and Richard S. Zemel. Learning latent subspaces in variational autoencoders. In *NeurIPS*, 2018. 3, 8

[19] Hieu Le, Bento Goncalves, Dimitris Samaras, and Heather Lynch. Weakly labeling the antarctic: The penguin colony case. In *CVPR Workshops*, June 2019. 1

[20] Hieu Le, Dimitris Samaras, and Heather J. Lynch. A convolutional neural network architecture designed for the automated survey of seabird colonies, 2022. 1

[21] Hieu Le, Chen-Ping Yu, Gregory Zelinsky, and Dimitris Samaras. Co-localization with category-consistent features and geodesic distance propagation. In *ICCV Workshop*, 2017. 2

[22] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. 2021. 6

[23] Yiting Li, Haiyue Zhu, Yu Cheng, Wenxin Wang, Chek Sing Teo, Cheng Xiang, Prahlad Vadakkepat, and Tong Heng Lee. Few-shot object detection via classification refinement and distractor retreatment. In *CVPR*, June 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[25] Jiawei Ma, Guangxing Han, Shiyuan Huang, Yuncong Yang, and Shih-Fu Chang. Few-shot end-to-end object detection via constantly concentrated encoding across heads. In *ECCV*, 2022. 2, 6

[26] 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. 2, 6

[27] Anay Majee, Kshitij Agrawal, and A. Subramanian. Few-shot learning for road object detection. *ArXiv*, abs/2101.12543, 2021. 1

[28] Anay Majee, A. Subramanian, and Kshitij Agrawal. Meta guided metric learner for overcoming class confusion in few-shot road object detection. *ArXiv*, abs/2110.15074, 2021. 1

[29] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 5, 6, 7

[30] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41, 1992. 4

[31] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, 2020. 1

[32] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, June 2020. 2

[33] 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, 2, 3, 5, 6, 7, 8

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 5, 6, 7

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 2015. 1, 5

[36] Mahdi Rezaei and Mahsa Shahidi. Zero-shot learning and its applications from autonomous vehicles to covid-19 diagnosis: A review. *Intelligence-Based Medicine*, 3:100005 – 100005, 2020. 1

[37] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Sharath Pankanti, Rogério Schmidt Feris, Abhishek Kumar, Raja Giryes, and Alexander M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *CVPR*, 2019. 2

[38] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek F. Abdelzaher. Controlvae: Controllable variational autoencoder. In *ICML*, 2020. 3

[39] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 4

[40] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, 2021. 1, 2, 3, 6

[41] Wenji Wang, Qing Xia, Zhiqiang Hu, Zhennan Yan, Zhuowei Li, Yang Wu, Ning Huang, Yue Gao, Dimitris N. Metaxas, and Shaoting Zhang. Few-shot learning by a cascaded framework with shape-constrained pseudo label assessment for whole heart segmentation. *IEEE Transactions on Medical Imaging*, 40:2629–2641, 2021. 1

[42] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *ArXiv*, abs/2003.06957, 2020. 3, 5, 6

[43] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *ICCV*, 2019. 2

[44] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1

[45] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. *ICCV*, pages 9547–9556, 2021. 2

[46] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4178–4193, 2022. 2

[47] Aming Wu, Suqi Zhao, Cheng Deng, and Wei Liu. Generalized and discriminative few-shot object detection via svd-dictionary enhancement. In *NeurIPS*, 2021. 2

[48] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. *ArXiv*, abs/2007.09384, 2020. 3, 6

[49] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 3

[50] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, 2020. 2, 6

[51] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *CVPR*, 2022. 3, 4

[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. 2

[53] Yukuan Yang, Fangyun Wei, Miaojing Shi, and Guoqi Li. Restoring negative information in few-shot object detection. *ArXiv*, abs/2010.11714, 2020. 2

[54] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *AAAI*, 2020. 2

[55] Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. *CVPR*, pages 13003–13012, 2021. 1, 3, 5, 6

[56] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-stable few-shot object detection. In *CVPR*, 2021. 2, 3, 6

[57] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, 2020. 1