GENERATING FEATURES WITH INCREASED CROP-RELATED DIVERSITY FOR FEW-SHOT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

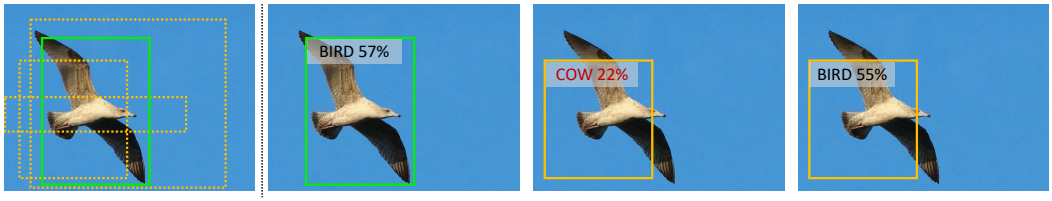
Two-stage object detectors generate object proposals and classify them to detect objects in images. These proposals often do not perfectly contain the objects but overlap with them in many possible ways, exhibiting great variability induced by different object scales, object positions (*w.r.t.* the boxes), object parts, and backgrounds. Training a robust classifier against this 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 variability. The main idea is to transform the latent space such that latent codes with different norms represent different crop-related variations. This allows us to generate features with increased crop-related diversity via 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 crop-related variation. 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. Our experimental results show that our generated features consistently improve state-of-the-art few-shot object detection methods on PASCAL VOC and 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 an 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 (Ouyang et al., 2020; Wang et al., 2021), rare animal species (Le et al., 2022; Welinder et al., 2010), satellite images (Borowicz et al., 2019; Le et al., 2019), or failure cases in autonomous driving systems (Rezaei & Shahidi, 2020; Majee et al., 2021a;b).

State-of-the-art FSOD methods are built on top of a two-stage framework (Ren et al., 2015), 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 (Sun et al., 2021). To mitigate this issue, a natural approach is to generate additional features for novel classes (Zhang & Wang, 2021; Zhu et al., 2020; Hayat et al., 2020). For example, Zhang & Wang (2021) assume that intra-class variation can be shared across categories. Hence, they 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. (2020) propose to synthesize visual features for unseen objects and incorporate them into unseen object detection. Hayat et al. (2020) propose to generate unseen features for novel classes based on semantic embedding using an adversarial framework.

Although much progress has been made, how to address the lack of data diversity is still an open question in FSOD. In this paper, we discuss a specific type of data variability that greatly affects the accuracy of FSOD algorithms. We observe that in FSOD, the classifier is not robust in classifying



(a) Different crops of an object instance | (b) DeFRCN (Qiao et al., 2021) | (c) DeFRCN (Qiao et al., 2021) | (d) Our proposed method

Figure 1: **Robustness to crop-related variability**. The region proposal network often outputs many object crops overlapping with the object instance in various way (a). The classifier head of the state-of-the-art FSOD method (Qiao et al., 2021) classifies correctly a simple crop of the bird (b) but mis-classifies a harder crop (c). Our method focuses on generating features with increased crop-related variability to use them as additional training data to improve the classifier (d).

different image crops (Figure 1a) of the same object instance. As can be seen from Figure 1, the state-of-the-art FSOD method, DeFRCN (Qiao et al., 2021), classifies correctly a simple crop of class "*bird*" (1b) but mis-classifies a harder case as "*cow*" (1c) when some parts of the bird are missing. In fact, our analysis shows that the performance of the method on those hard cases is significantly worse than on easy cases (Section 5.3). A potential reason is that the features of different crops of the same object instance can be vastly different due to changes in object scales, object parts included in the crops, object positions within the crops, and backgrounds. However, training a classifier that is robust to this crop-related variability is challenging due to limited training data in few-shot settings (Zhang & Wang, 2021). To the best of our knowledge, no previous FSOD method focuses on addressing this issue.

In this paper, we mitigate this issue by generating additional data with increased crop-related diversity for novel classes. Note that previous few-shot learning methods mainly focus on increasing intra-class variance in general while our goal is to enforce the diversity *w.r.t.* a specific property in generated samples. It is challenging to capture and model a specific source of variability since there are inherently multiple factors of variation in the data (Mathieu et al., 2016). To accomplish this goal, we propose a novel generative model in which crop-related variation can easily be manipulated from the learned latent representations. The key idea is to rescale each latent code such that its norm (magnitude) directly represents the crop-related variation in the input. As different norms represent different variations, we can generate features with increased crop-related diversity by simply changing the latent norm.

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 of the latent code with how the object proposal overlaps with the ground-truth bounding box. 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 partially characterizes the overlap: 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. In essence, we use these IoU scores as proxies to represent different cropping variations. By directly encoding the crop-related variation into the norm of the latent code, we can easily manipulate the crop-related variation of the generated samples.

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 class. In our experiments, we use our generated samples to fine-tune the baseline few-shot object detector - DeFRCN (Qiao et al., 2021). Somewhat 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 (Everingham et al., 2009). Norm-VAE, our proposed VAE, further improves 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 (Qiao et al., 2021) for both PASCAL VOC and COCO (Lin et al., 2014) datasets.

Our main contributions can be summarized as follows:

- We show that lack of crop-related variability 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 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 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 (Kaul et al., 2022; Han et al., 2021; Perez-Rua et al., 2020; Fan et al., 2020b; Zhu et al., 2021; Sun et al., 2021; Li et al., 2021b; Fan et al., 2021; Wu et al., 2021b;a; 2022; Kaul et al., 2022; Han et al., 2022) 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 (Kang et al., 2019; Yan et al., 2019; Xiao & Marlet, 2020; Schwartz et al., 2019; Yang et al., 2020a;b; Fan et al., 2020a; Wang et al., 2019). Meta-learning based approaches introduce a meta-learner to acquire meta-knowledge that can be then transferred to novel classes. Kang et al. (2019) 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. Wang et al. (2019) propose 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 (Wang et al., 2020; Wu et al., 2020; Sun et al., 2021; Cao et al., 2021; Qiao et al., 2021). Wang et al. (2020) 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 (Sun et al., 2021) 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 (Zhang & Wang, 2021). 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, (Xu & Le, 2022) 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. (Xian et al., 2019) 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, (Zhang & Wang, 2021) propose to transfer the shared modes of within-class variation from base classes to novel classes to hallucinate new samples. (Zhu et al., 2021) propose to synthesize visual features for unseen objects from semantic information and augment existing training algorithms to incorporate unseen object detection. Recently, (Huang et al., 2022) 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 (Higgins et al., 2017; Klys et al., 2018; Via et al., 2021; Shao et al., 2020). β-VAE (Higgins et al., 2017) 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 (Shao et al., 2020) 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 (Klys et al., 2018) and PCVAE (Via et al., 2021). CSVAE (Klys et al., 2018) learns latent dimensions associated with binary properties. The learned latent subspace can easily be inspected and independently manipulated. PCVAE (Via et al., 2021) uses 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 one variational factor into the norm of the latent code. Experiments show that our strategy outperforms other VAE architectures, albeit 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, an region proposal network (RPN) takes the output feature maps from a backbone feature extractor as inputs and generates region proposals. An 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 (Qiao et al., 2021). In general, our method can be built on top of any two-stage object detection methods.

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.

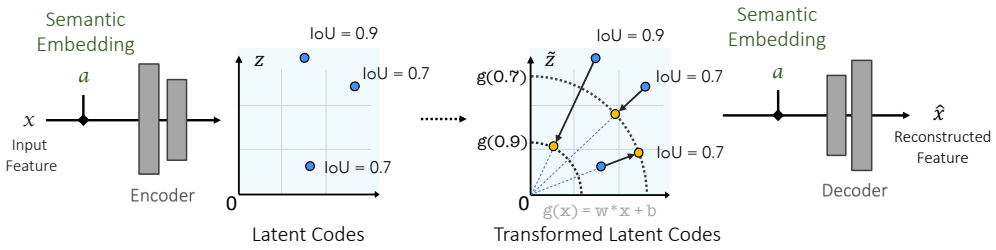


Figure 2: **Norm-VAE for modelling crop-related variations.** The original latent code z is rescaled to $\hat{z}$ such that the norm of $\hat{z}$ linearly correlates with the IoU score of the input crop (*w.r.t.* the ground truth box). The original latent codes are colored in **blue** while the rescaled 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.

3.2.1 NORM-VAE FOR FEATURE GENERATION

We develop our feature generator based on a conditional VAE architecture (Sohn et al., 2015). 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 (Miller, 1992; Radford et al., 2021). 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 (Xu & Le, 2022). 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:

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

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:

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

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 $\mathbb{G}^y$:

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

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 Zhang & Wang (2021) 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) (Everingham et al., 2009) and MS COCO datasets (Lin et al., 2014). For fair comparison, we follow the data split construction and evaluation protocol used in previous works (Kang et al., 2019). The PASCAL VOC dataset contains 20 categories. We use the same 3 base/novel splits with TFA (Wang et al., 2020) 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

Our method can be built on top of many existing few-shot object detection methods. In our experiments, we use the pre-trained Faster-RCNN (Ren et al., 2015) with ResNet-101 (He et al., 2016) and Feature Pyramid Network (Lin et al., 2017) following previous work DeFRCN (Qiao et al., 2021). 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 (Radford et al., 2021) model in all main experiments. An additional experiment using Word2Vec (Mikolov et al., 2013) embeddings is reported in Section 5.2. After the VAE is trained on the base set with various augmented object boxes (Appendix C), we use the trained decoder to generate $N = 30$ features (Appendix D) 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 to the norm of the new latent code. We empirically choose $g(\cdot)$ such that $g(1) = \sqrt{512}$ and $g(0.5) = 5 * \sqrt{512}$[1]. Note that $g(\cdot)$ enforces an inverse correlation: the features with low IoU scores, which are considerably hard samples, are mapped to a higher norm and placed further away from the origin (Meng et al., 2021). We provide further analyses on the choice of $g(\cdot)$ in Appendix E. 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.

[1] 512 is the dimension of the latent code

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 (Wang et al., 2020)	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 (Wang et al., 2020)	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 (Wu et al., 2020)	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 (Xiao & Marlet, 2020)	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 (Sun et al., 2021)	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 (Li et al., 2021a)	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 (Zhu et al., 2021)	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. (Zhang & Wang, 2021)	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
FADI (Cao et al., 2021)	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
Pseudo-Labelling (Kaul et al., 2022)	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
FCT (Han et al., 2022)	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
DeFRCN (Qiao et al., 2021)	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 (Wang et al., 2020)	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 (Wang et al., 2020)	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 (Wu et al., 2020)	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 (Cao et al., 2021)	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 (Han et al., 2022)	-	7.9	-	-	17.1	21.4	-	7.9	-	-	17.0	22.1
Pseudo-Labeling (Kaul et al., 2022) †	-	-	-	-	17.8	**24.5**	-	-	-	-	**17.8**	**25.0**
DeFRCN (Qiao et al., 2021)	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 COCO dataset**. Our approach outperforms baseline methods in most cases, especially in low-shot settings (K<10). † applies mosaic data augmentation introduced in (Bochkovskiy et al., 2020) during fine-tuning. Best performance in bold.

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 Appendix F.

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<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 (Bochkovskiy et al., 2020) 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 3, using the vanilla VAE with more training data does not bring performance improvement compared to the VAE model trained with the 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 (Radford et al., 2021) features in our main experiments. In Table 4, we compare this model with another model trained with Word2Vec (Mikolov et al., 2013) on PASCAL VOC dataset.

	Data	1-shot	2-shot	3-shot
DeFRCN (Qiao et al., 2021)	-	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 3: **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.

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.

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 (Qiao et al., 2021)	-	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	Word2Vec	**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	CLIP	**62.1**	**64.9**	**67.8**	**39.9**	**46.8**	**54.4**	**58.2**	**60.3**	**61.0**

Table 4: **FSOD Performance of VAE models trained with different class semantic embeddings**. CLIP (Radford et al., 2021) is trained with 400M pairs (image and its text title) collected from the web while Word2Vec (Mikolov et al., 2013) is trained with only text data.

5.3 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 (see Appendix A and B for further analyses and results). However, the performance is still far from perfect. Addressing these challenging cases is a fruitful venue for future FSOD work.

Method	1-shot	2-shot	3-shot
DeFRCN(Qiao et al., 2021)	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.

5.4 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).

Features	1-shot		2-shot		3-shot		5-shot	
	nAP50	nAP75	nAP50	nAP75	nAP50	nAP75	nAP50	nAP75
Low-IoU	**60.9**	30.5	**63.7**	40.6	**66.6**	40.7	**68.9**	41.2
High-IoU	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" features has better nAP50 scores while the "High-IoU" model has better nAP75 scores.

5.5 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 (Higgins et al., 2017) and CSVAE (Klys et al., 2018). 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 fine-tune 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 Qiao et al. (2021)	56.3	60.3	62.0
β-VAE(Higgins et al., 2017)	61.3	64.0	67.3
CSVAE(Klys et al., 2018)	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 COCO datasets. Our proposed VAE model is simple, easy to implement, and allows modifying a specific property 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. 3) IoU score might not be a sufficient proxy to represent crop-related variation. A single IoU value might represent multiple variational factors, which is not ideal. Disentangling those factors and representing them in the embedding space to effectively diversify generated data can be a natural extension of our work.

REFERENCES

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

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.

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.

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: 303–338, 2009.

Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4012–4021, 2020a.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020b.

Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4527–4536, June 2021.

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

Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5321–5330, 2022.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust region feature synthesizer for zero-shot object detection. 2022.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8419–8428, 2019.

Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few-shot object detection method. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

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

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

Hieu Le, Dimitris Samaras, and Heather J. Lynch. A convolutional neural network architecture designed for the automated survey of seabird colonies. *Remote Sensing in Ecology and Conservation*, 8(2):251–262, 2022. doi: https://doi.org/10.1002/rse2.240. URL https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.240.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7359–7368, 2021a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15395–15403, June 2021b.

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.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

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

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, 2021b.

Michaël Mathieu, Junbo Jake Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *NIPS*, 2016.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14220–14229, 2021.

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

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

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.

Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8661–8670, 2021.

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.

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*, 39:1137–1149, 2015.

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.

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. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5192–5201, 2019.

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.

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

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7348–7358, 2021.

Coder Via, Invertible Mutual Dependence, Xiaojie Guo, Yuanqi Du, and Liang Zhao. Property controllable variational autoen-. 2021.

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.

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

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9924–9933, 2019.

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.

Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9547–9556, 2021a.

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

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.

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

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10267–10276, 2019.

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

Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *CVPR*, 2022.

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. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9576–9585, 2019.

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

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

Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13003–13012, 2021.

Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-stable few-shot object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8778–8787, 2021.

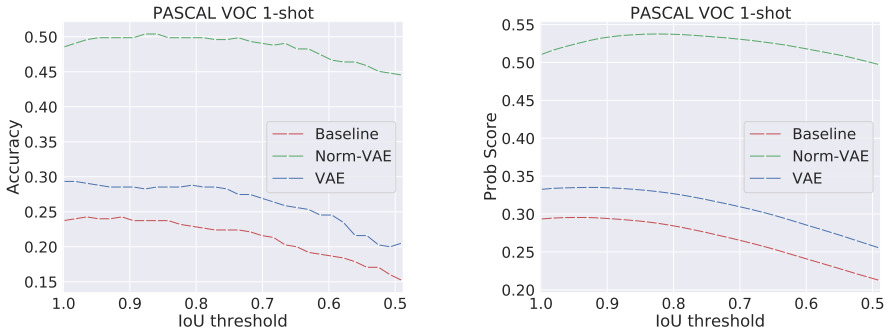
Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11690–11699, 2020.

A 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 DeFRCNQiao et al. (2021), "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.

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.

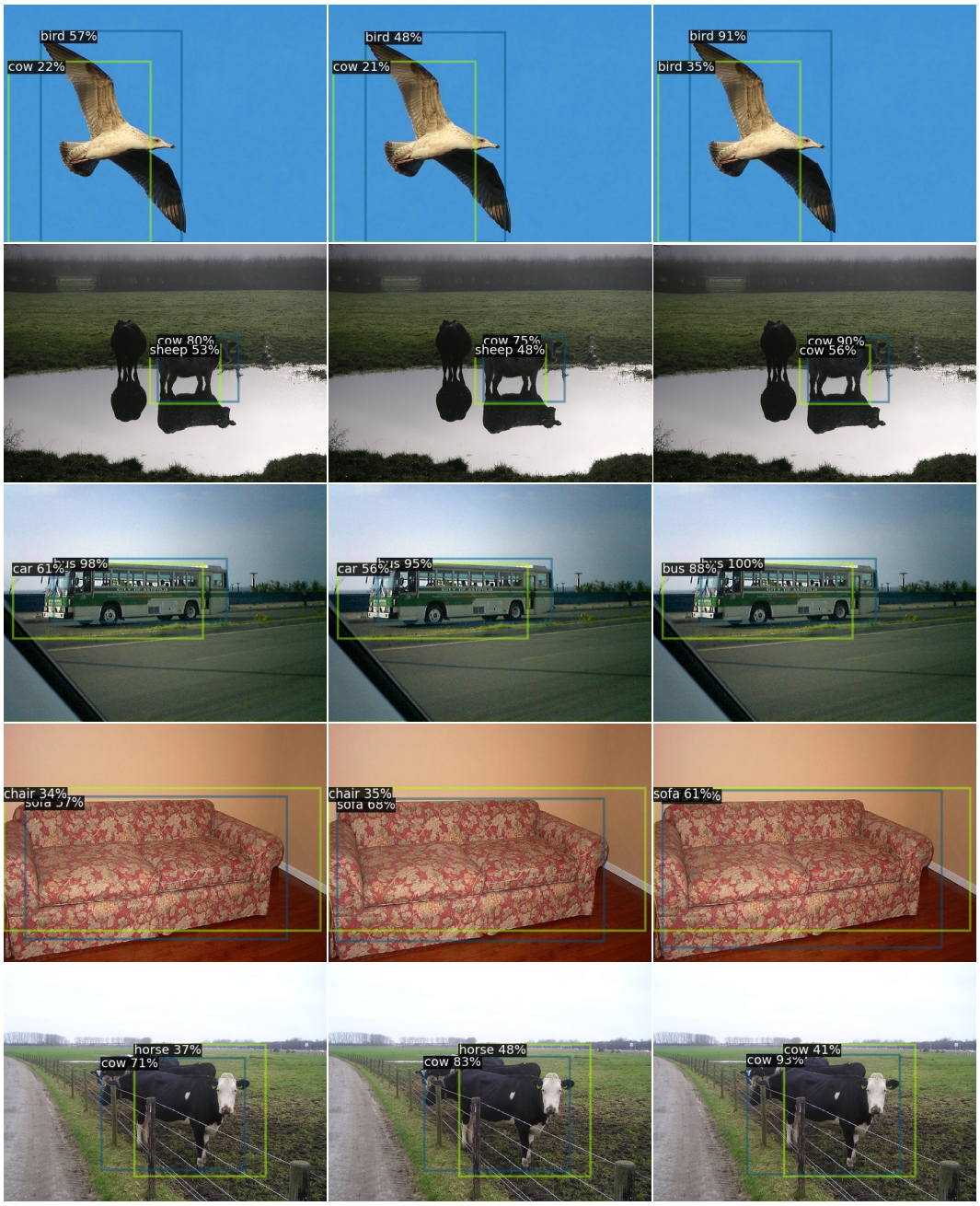


(a) Classification accuracy (b) Class probability score

Figure 3: **Classification accuracy and probability score of the object detector on the augmented box**. We compare between the baseline DeFRCN Qiao et al. (2021), 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.

B Detection Results for Inaccurate Bounding Boxes

In this section, we provide qualitative visualizations of the detected objects of the 1-shot model on PASCAL VOC Novel Split 1. As shown in Figure 4, for each input image, the blue box is the original prediction result from the object detector. We then randomly create an augmented bounding box based on the ground-truth bounding box and input the augmented box to the classifier of the object detector. The prediction result on the augmented box is denoted as the yellow box. For the examples shown in the figure, the baseline DeFRCN model (Qiao et al., 2021) and the model trained with features from a vanilla VAE predict the class labels correctly on the original input boxes while both fail on the augmented boxes. By contrast, the model trained with features from Norm-VAE can classify both the original box and the augmented box correctly. As can be seen, crop-related variation is crucial for object detection and our method can enhance the object detector's robustness against the variation successfully.



DeFRCN (Qiao et al., 2021) Vanilla-VAE Norm-VAE

Figure 4: **Qualitative visualizations of the detected objects on PASCAL Novel Split 1**. "Vanilla-VAE" denotes the model trained with features generated from a vanilla VAE and "Norm-VAE" denotes the model trained with features generated from Norm-VAE. The blue box is the detector's prediction on the original image and the yellow box is the prediction on the augmented box. Our proposed Norm-VAE can generate features that enhance the model's robustness against crop-related variation.

C Details on Generating Augmented Training Data

We extract the image features from image crops from the base classes and use them to train a feature generator to generate features for the novel classes. Specifically, we apply the RoI head

feature extractor on the ground-truth bounding box b_i from the base classes to get the RoI feature f_i. To enrich the diversity of the RoI feature, we randomly create N additional augmented bounding boxes by randomly moving the starting point and the ending point of the original box, annotated as $\{b_i^1, b_i^2, ...b_i^N\}$. These augmented bounding boxes overlap the ground-truth bounding box differently and have different IoU scores. With a set of augmented bounding boxes $\{b_i^1, b_i^2, ...b_i^N\}$, we extract the corresponding RoI features $\{f_i^1, f_i^2, ...f_i^N\}$ and use them to train our VAE model.

D Number of Generated Samples

In our main experiment, we generate 30 samples per class and use them together with the original few-shot samples to fine-tune the object detector. In this section, we investigate the impact of the number of the generated samples. Table 8 shows the AP50 on PASCAL VOC Novel Split 1 with different numbers of generated features under 1-shot, 2-shot and 3-shot settings. As the number of generated samples increases, the performance gradually improves and then plateaus and drops slightly (less than 0.5% decrease in performance).

# Generated Features	0	5	10	15	20	25	30	35
1-shot	56.3	60.5	61.6	61.8	62.0	61.9	**62.1**	62.0
2-shot	60.3	62.0	63.7	63.6	63.6	64.1	**64.9**	64.5
3-shot	62.0	65.6	67.0	67.2	67.2	**67.8**	**67.8**	67.3

Table 8: **Impact of the number of the generated samples under PASCAL VOC Novel Split 1**. As the number of generated samples increases, the performance gradually improves and then saturates and drops slightly.

E Mapping Function Analyses

We use a simple pre-defined linear function $g(x) = w \times x + b$ to map from an IoU score x to the new norm of a latent code. Here we only consider proposals with IoU scores ranging from 0.5 to 1. The coefficients of the linear function can be inferred by defining the value range of the rescaled latent norm. We conduct experiments with different ranges and the results are shown in Table 9. Note that here $\sqrt{512}$ is a scaling constant that corresponds to the number of dimension ($N = 512$) of the latent space. As can be seen from the table, we observe better performance when the IoU score inversely correlates the latent norm. In this case, a proposal with low IoU score has a higher latent norm and is placed further away from the origin. A possible reason is that features of hard instances often exhibit higher variance. Thus, it is more optimal to use latent codes with larger norms to represent them (Meng et al., 2021).

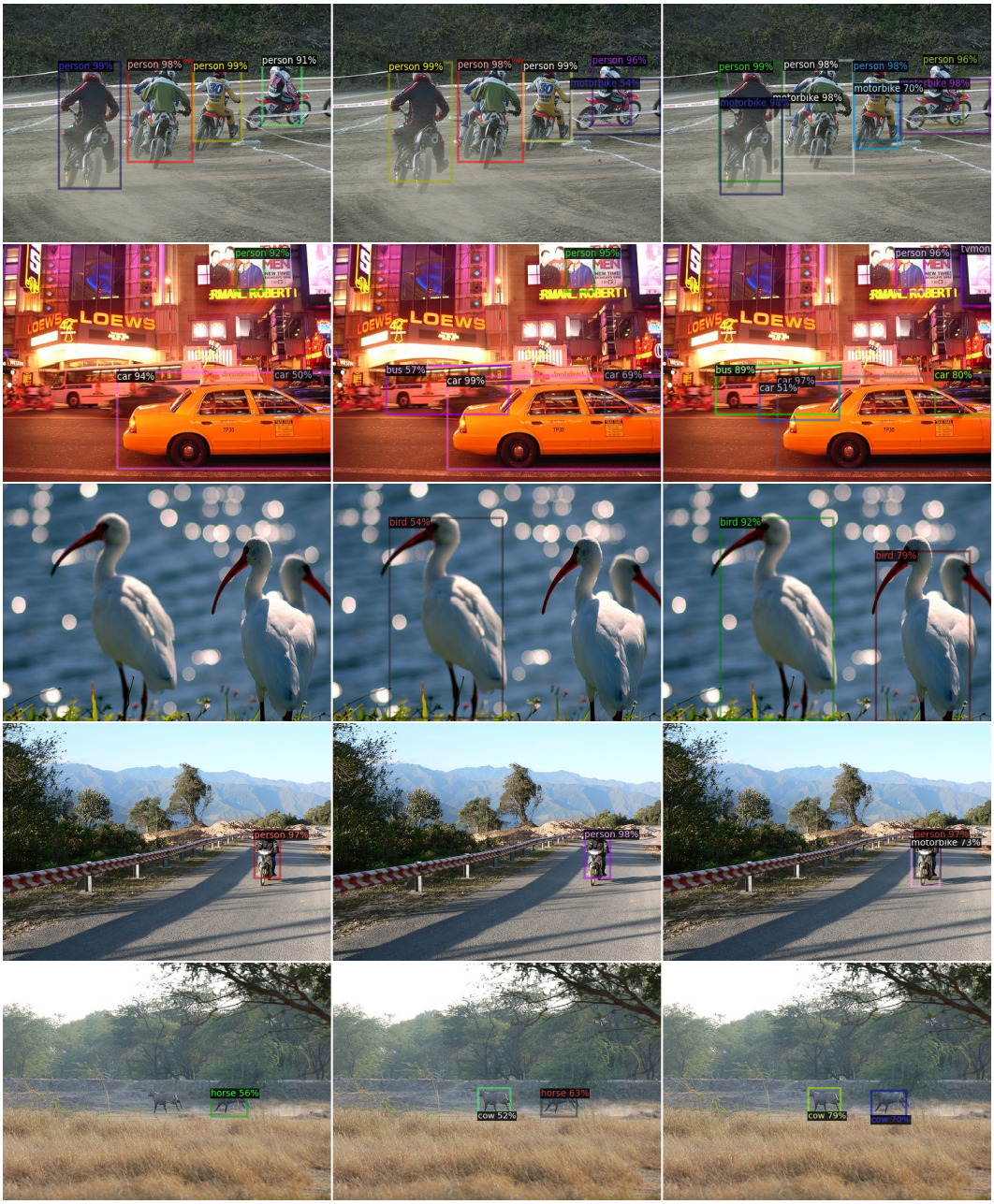
	$g(1)$	$g(0.5)$	AP50
Inverse Correlation	$1\times\sqrt{512}$	$2\times\sqrt{512}$	61.6
	$1\times\sqrt{512}$	$5\times\sqrt{512}$	**62.1**
	$1\times\sqrt{512}$	$10\times\sqrt{512}$	61.8
Correlation	$2\times\sqrt{512}$	$1\times\sqrt{512}$	60.6
	$5\times\sqrt{512}$	$1\times\sqrt{512}$	61.3
	$10\times\sqrt{512}$	$1\times\sqrt{512}$	60.6

Table 9: **Performance with different configurations of the mapping function**. We conduct experiments using different coefficients for function $g(\cdot)$, which defines the value range of the new norm of the latent code.

F Visualization of the Detection Results on PASCAL VOC dataset

We visualize the detection results of DeFRCN Qiao et al. (2021) and our proposed method in Figure 5. "Vanilla-VAE" denotes the model trained with features generated from a vanilla VAE and "Norm-VAE" denotes the model trained with features generated from Norm-VAE.

In all cases, the model trained with additional features performs better than DeFRCN. In the third row, DeFRCN fails to recognize both the two instances of the "*bird*" class while both Vanilla-VAE and Norm-VAE recognize them. It can be seen that with additional data from Norm-VAE, the FSOD model can recognize objects that are undetected with the model trained with just the original training data. The Norm-VAE model is generally more robust in recognizing objects. It works well even when they are cropped (2nd row) or small (two bottom rows).



DeFRCN Qiao et al. (2021) Vanilla-VAE Norm-VAE

Figure 5: **Visualization of the detection results on PASCAL VOC dataset.** "Vanilla-VAE" denotes the model trained with features generated from a vanilla VAE and "Norm-VAE" denotes the model trained with features generated from Norm-VAE. In all cases, the model trained with additional features performs better than DeFRCN. The Norm-VAE model works well even when the objects are partially cropped (2nd row) or small (two bottom rows). Detection score threshold is 0.5. Please view in magnification for cases with small objects.