Discriminating Known from Unknown Objects via Structure-Enhanced Recurrent Variational AutoEncoder

Aming Wu, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an, China
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

Abstract

Discriminating known from unknown objects is an important essential ability for human beings. To simulate this ability, a task of unsupervised out-of-distribution object detection (OOD-OD) is proposed to detect the objects that are never-seen-before during model training, which is beneficial for promoting the safe deployment of object detectors. Due to lacking unknown data for supervision, for this task, the main challenge lies in how to leverage the known in-distribution (ID) data to improve the detector's discrimination ability. In this paper, we first propose a method of Structure-Enhanced Recurrent Variational AutoEncoder (SR-VAE), which mainly consists of two dedicated recurrent VAE branches. Specifically, to boost the performance of object localization, we explore utilizing the classical Laplacian of Gaussian (LoG) operator to enhance the structure information in the extracted low-level features. Meanwhile, we design a VAE branch that recurrently generates the augmentation of the classification features to strengthen the discrimination ability of the object classifier. Finally, to alleviate the impact of lacking unknown data, another cycle-consistent conditional VAE branch is proposed to synthesize virtual OOD features that deviate from the distribution of ID features, which improves the capability of distinguishing OOD objects. In the experiments, our method is evaluated on OOD-OD, open-vocabulary detection, and incremental object detection. The significant performance gains over baselines show the superiorities of our method. The code will be released at https://github.com/AmingWu/SR-VAE.

1. Introduction

Recent years have witnessed the rapid development of deep learning based object detection [5,12,34,36,41], which often follows a close-set assumption that the training and testing processes share the same category space. However, the practical scenario is open and filled with unknown objects, e.g., in Fig. 1, an autonomous vehicle may encounter an unseen camel, presenting significant challenges for close-set assumption based detectors. To promote the safe application of detectors, a task of unsupervised out-of-distribution object detection (OOD-OD) [7] is recently proposed, which aims to detect the objects never-seen-before during training without accessing any auxiliary data.

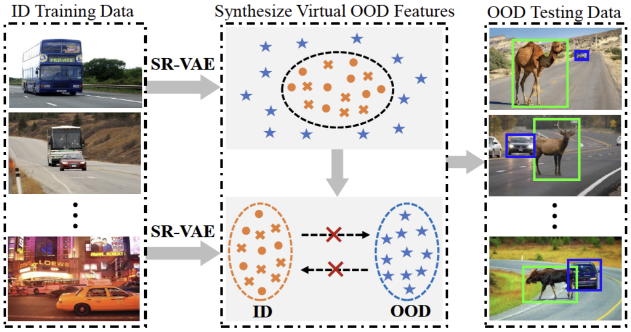


Figure 1. Discriminating known from unknown objects (as shown in green boxes) by synthesizing virtual OOD features. For OOD-OD, to alleviate the impact of lacking unknown data, we present an SR-VAE method to constrain the synthesized features (as shown in blue stars) to deviate from the distribution of ID features (as shown in orange). Meanwhile, we consider enhancing the discrimination of the classifier to reduce the risk of misclassifying the ID objects into the OOD category. Through these operations, the ability of distinguishing OOD objects could be improved significantly.

Towards unsupervised OOD-OD, since there is no auxiliary data available for supervision, leveraging the known in-distribution (ID) data to enhance the detector's discrimination ability becomes the critical challenge. One feasible solution is to synthesize a series of virtual OOD features [7, 35] based on the ID data, which is beneficial for promoting the object detector to learn a clear decision boundary between ID and OOD objects. To this end, the work [7] attempts to synthesize virtual features from the low-likelihood region of the estimated class-conditional distribution. However, this method requires a large number of objects for each category to estimate the distribution, limiting its application to the case of few samples.

As shown in Fig. 1, in this paper, we consider improv-

[*] Corresponding author.

ing the performance of OOD object detection from two perspectives: One is to strengthen the discrimination capability of the object classifier for the known ID objects, which is conducive to reduce the risk of misclassifying the ID objects into the OOD category. Another is to synthesize the virtual OOD features that significantly deviate from the distribution of the ID features, which is instrumental in boosting the performance of distinguishing OOD objects from ID objects. To attain these two goals, we explore exploiting Variational AutoEncoder (VAE) [15, 20] to separately generate the augmented ID features and virtual OOD features.

Specifically, an approach of Structure-Enhanced Recurrent Variational AutoEncoder (SR-VAE) is proposed, which mainly consists of two dedicated recurrent VAE branches. In general, an object detector should first localize objects. Then, an object classifier is used to discriminate these objects [12, 36]. To improve the localization performance, enhancing the object-related information in the extracted features is meaningful. To this end, after extracting the low-level features of an input image, we present to utilize the classical Laplacian of Gaussian (LoG) operator [23] to obtain structure-related information, which is used to fuse into the existing features to strengthen the localization ability.

Next, in order to enhance the discrimination ability, inspired by Invariant Risk Minimization [1], we explore constructing a set of diverse environments to intensify the variance of the input classification features. Concretely, a VAE module [17, 20, 43] is exploited to recurrently output multiple augmented features of the classification features, i.e., the current output is taken as the input of the next iteration. Since the input of each iteration is different, by means of the variation operation, the diversity of the output features could be enlarged. Then, the discrimination ability is improved by minimizing the prediction discrepancy between the augmented features and the classification features.

Finally, to alleviate the impact of lacking unknown data, we present a cycle-consistent conditional VAE [40] to synthesize virtual OOD features in the absence of paired supervision samples. Concretely, to ameliorate the synthesized features to deviate from the distribution of ID features, we first insert label information in the latent space to force a deterministic constrained representation. Meanwhile, by maximizing the discrepancy between the synthesized features and the input features, the synthesized features could be facilitated to contain plentiful OOD-relevant information, which enhances the ability of distinguishing OOD objects. In the experiments, our method is separately evaluated on three different tasks. Extensive experimental results on multiple datasets demonstrate the superiorities of our method.

The contributions are summarized as follows:

(1) For unsupervised OOD-OD, we observe that using the classical LoG operator could effectively enhance object-related information in the extracted low-level features.

(2) To reduce the risk of misclassifying ID objects into the OOD category, we design a dedicated recurrent VAE to generate diverse augmented features of the input classification features, which is beneficial for improving the discrimination ability of the object classifier.

(3) To alleviate the impact of lacking unknown data for supervision, we present a cycle-consistent conditional VAE to synthesize virtual OOD features, which is instrumental in distinguishing OOD objects from ID objects.

(4) In the experiments, our method is evaluated on OOD-OD [7], open-vocabulary detection [33, 49], and incremental object detection [22, 39]. Particularly, for OpenImages dataset [24], compared with the baseline method [7], our method significantly reduces FPR95 by around **13.73%**.

2. Related Work

OOD Detection. In order to promote the safe application of models in real scenarios, OOD detection [4, 14, 18] has recently attracted much attention, whose goal is to distinguish OOD data from ID data. Most methods [27, 31, 46, 52] focus on OOD image classification and explore a proper regularization-based method for this task. Particularly, Bendale et al. [2] developed the OpenMax score based on the extreme value theory. Liu et al. [30] proposed to utilize the energy score to discriminate OOD data from ID data. Zhou [52] rethought the commonly used reconstruction autoencoder for OOD detection. Though these methods have been demonstrated to be effective, since object detection involves object localization and classification, these methods could not directly apply to OOD-OD.

Recently, unsupervised OOD-OD [7] is proposed to localize and recognize the objects never-seen-before during training. For this task, Du et al. [7] proposed to use large-scale samples to estimate the distribution of each ID category, which is used to synthesize virtual OOD features. However, calculating accurate distribution estimation may limit its application in real scenarios. Harakeh et al. [11] designed an uncertainty estimation method for the localization branch, which could not well address OOD-OD that involves localization and classification. Besides, the work [6] proposed to use auxiliary video datasets to learn unknown-aware knowledge to effectively improve the performance of distinguishing OOD objects, which could not be used for unsupervised OOD-OD. In this paper, we explore employing the idea of VAE [15, 20] to synthesize virtual OOD features that significantly deviate from the distribution of ID features. And these virtual features are used to strengthen the discrimination ability. Experimental results on three different tasks show the effectiveness of our method.

Variational AutoEncoder. The goal of VAE [15, 20, 21, 40] is to map the input data to a multivariate latent distribution, which efficiently addresses the generation problem for high-dimensional data. Most existing methods [15, 17, 44]

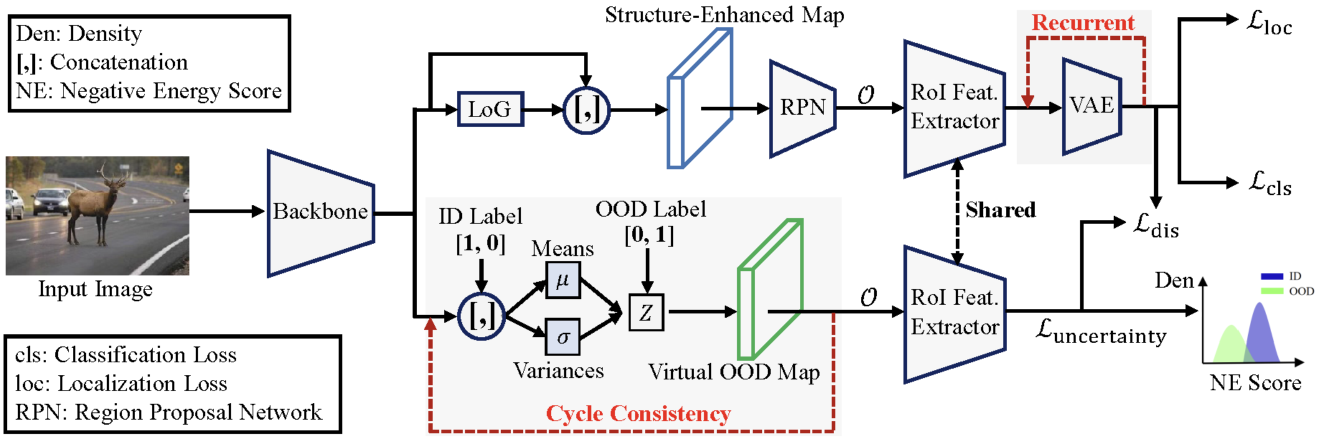


Figure 2. Structure-Enhanced Recurrent Variational AutoEncoder for Unsupervised OOD-OD. To improve localization performance, we employ the LoG operator to enhance the structure-related information in the features extracted by the backbone network. Meanwhile, we design a VAE module to recurrently generate diverse augmented features of the input classification features, which is beneficial for strengthening the discrimination ability of the object classifier. Finally, to alleviate the impact of lacking unknown data, we propose a cycle-consistent conditional VAE to synthesize virtual OOD features that deviate from the distribution of the ID features significantly, which boosts the performance of distinguishing OOD objects.

attempt to adjust the latent representations to generate clear images containing rich content. Particularly, β-VAE [15] is an implementation with a weighted Kullback-Leibler (KL) divergence term to automatically discover disentangled representations without supervision. Conditional VAE [40] inserts label information in the latent space to force a deterministic constrained representation. Different from the above works focusing on pixel-level generation, we explore using VAE to generate virtual OOD features. Experimental results on OOD-OD show the advantages of our method.

3. Structure-Enhanced Recurrent VAE

In this paper, we follow the settings introduced in the work [7]. During training, we can only access the ID data. During inference, the object detector should own the capability of distinguishing ID objects from OOD objects.

3.1. Structure Enhancement via LoG Operator

In general, an object detector should first localize objects accurately and then distinguish them as ID categories or OOD categories. To this end, it is important to enhance object-related information (e.g., object structure information). In the field of image processing, LoG operator [23] is a popular algorithm for edge detection. We explore performing the LoG operation on the extracted low-level features to strengthen the structure-relevant information.

Concretely, as shown in Fig. 2, we follow the baseline work [7] and adopt the widely used object detector, i.e., Faster R-CNN [12, 36], as the basic detection model. Given an input image, we first employ the backbone network, e.g., ResNet [13], to extract the corresponding feature map $F \in \mathbb{R}^{w \times h \times c}$, where w, h, and c separately denote width, height, and channel number. To obtain rich structure information, a 9×9 LoG kernel $\mathcal{G}$ is defined as follows:

$$\mathcal{G} = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 2 & 5 & 0 & -24 & -40 & -24 & 0 & 5 & 2 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \end{bmatrix}. \tag{1}$$

Next, $\mathcal{G}$ is used to perform a convolution operation on F to obtain the structure-enhanced map $E \in \mathbb{R}^{w \times h \times c}$. The processes are shown as follows:

$$\mathcal{E} = F * \mathcal{G}, \qquad E = \Psi([F, \mathcal{E}]), \tag{2}$$

where $\mathcal{E} \in \mathbb{R}^{w \times h \times c}$ is the convolutional result containing plentiful structure-related information. $\Psi(\cdot) \in \mathbb{R}^{1 \times 1 \times 2c \times c}$ represents one-layer convolution to transform the number of channels. By fusing the structure information into existing features, the object-related information in F could be enhanced effectively, which is instrumental in improving the performance of object localization.

3.2. Recurrent VAE for Improving Discrimination

To reduce the risk of misclassifying ID objects into the OOD category, we design a VAE module to recurrently generate diverse augmented features of the classification features, which enhances the discrimination ability.

Specifically, as shown in Fig. 2, the enhanced map E is taken as the input of the RPN module to obtain a set of object proposals $\mathcal{O}$. Based on $\mathcal{O}$, RoI-Alignment followed by RoI-Feature extraction [12] is performed on E to output

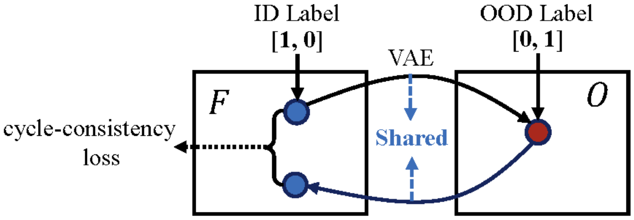


Figure 3. Cycle-consistent conditional VAE for synthesizing virtual OOD features. During the forward process, the OOD features $\boldsymbol{O}$ are generated based on the ID features F. And during the backward process, the features $\boldsymbol{F}$ are generated based on $\boldsymbol{O}$. By minimizing the cycle-consistency loss between F and $\boldsymbol{F}$, the synthesized features $\boldsymbol{O}$ could be promoted to contain plentiful information that deviates from the distribution of the ID features.

$P_{\text{in}} \in \mathbb{R}^{m\times n}$, where m and n separately denote the number of proposals and channels. Next, we first define two fully-connected networks (i.e., $\Phi_{\mu}(\cdot)$ and $\Phi_{\sigma}(\cdot)$) as the encoder to estimate the corresponding means and variances. Then, an encoding operation is performed on the input. Finally, a decoder $\Theta(\cdot)$ consisting of three fully-connected layers is leveraged to calculate the output. The overall processes at the t-th iteration are shown as follows:

$$\begin{aligned} \mu_t &= \Phi_{\mu}(H_{t-1}), & \sigma_t &= \Phi_{\sigma}(H_{t-1}), \\ h_t &= \mu_t + \epsilon \cdot \exp(\sigma_t), & H_t &= \Theta(h_t), \end{aligned} \tag{3}$$

where $H_{t-1} \in \mathbb{R}^{m\times n}$ and $H_t \in \mathbb{R}^{m\times n}$ respectively indicate the output of the previous iteration and that of the current iteration. And $H_0 = P_{\text{in}}$. $\mu_t \in \mathbb{R}^{m\times n}$ and $\sigma_t \in \mathbb{R}^{m\times n}$ are the estimated means and variances. ϵ denotes Gaussian noise sampled from $\mathcal{N}(0, I)$.

Through T iterations, we can obtain a set of augmented features $H = \{H_1, ..., H_T\}$. Finally, by minimizing the KL-divergence between the prediction probabilities from H_t and P_{in}, the discrimination ability of the object classifier could be enhanced effectively. During training, P_{in} is taken as the input of the object classifier and regressor to calculate the classification and localization losses. The joint training objective is shown as follows:

$$\mathcal{L}_{\text{in}} = \mathcal{L}_{\text{det}} + \alpha \cdot \frac{1}{T}\sum_{t=1}^{T} \text{KL}[p(H_t|H_{t-1}), p(P_{\text{in}})], \tag{4}$$

where $\mathcal{L}_{\text{det}}$ is the sum of the classification loss $\mathcal{L}_{\text{cls}}$ and the localization loss $\mathcal{L}_{\text{loc}}$, i.e., $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}}$. α is a hyper-parameter. In the experiments, α is set to 0.001. $p(\cdot)$ indicates the prediction probability of the object classifier.

3.3. Synthesizing Virtual OOD Features

Due to lacking unknown data for supervision, it is prone to misclassify OOD objects into ID categories. Hence, synthesizing virtual OOD data is important to distinguish OOD objects. A straightforward idea is to train generative models, e.g., GANs [9], to synthesize images, which is difficult to optimize [7]. Instead, we propose a cycle-consistent conditional VAE to synthesize virtual OOD features.

Particularly, as shown in Fig. 3, during the forward process, the extracted features F and corresponding ID label $[1,\ 0]$ are first concatenated as $\hat{F} \in \mathbb{R}^{w\times h\times(c+2)}$. Then, two convolutional networks W_{μ} and W_{σ} are defined to estimate the corresponding means and variances:

$$\mu_f = W_{\mu} * \hat{F}, \qquad \sigma_f = W_{\sigma} * \hat{F}, \tag{5}$$

where $\mu_f \in \mathbb{R}^{w\times h\times c}$ and $\sigma_f \in \mathbb{R}^{w\times h\times c}$. And an encoding operation is performed to obtain $Z \in \mathbb{R}^{w\times h\times c}$, i.e., $Z = \mu_f + \epsilon \cdot \exp(\sigma_f)$. Next, in order to generate virtual OOD features, Z and corresponding OOD label $[0,\ 1]$ are concatenated as $\hat{Z} \in \mathbb{R}^{w\times h\times(c+2)}$. A decoding network $\mathcal{D}$ consisting of three convolutional layers is performed on $\hat{Z}$ to obtain the virtual OOD map $\boldsymbol{O} \in \mathbb{R}^{w\times h\times c}$.

During the backward process, to alleviate the impact of lacking paired data for supervision, the concatenation result of $\boldsymbol{O}$ and label $[0,\ 1]$ is made convolution with W_{μ} and W_{σ}. Next, we still take the concatenation of the encoding result and label $[1,\ 0]$ as the input of the decoder $\mathcal{D}$ to obtain the output $\boldsymbol{F} \in \mathbb{R}^{w\times h\times c}$. The operations are the same as the forward process. Finally, we introduce a cycle-consistency loss to reduce the discrepancy between $\boldsymbol{F}$ and F, i.e., $\mathcal{L}_{\text{cycle}} = \frac{1}{wh}\sum|\boldsymbol{F} - F|$.

Moreover, to promote $\boldsymbol{O}$ to deviate from the distribution of F, as shown in Fig. 2, based on the object proposals $\mathcal{O}$, RoI-Alignment followed by RoI-Feature extraction is performed on $\boldsymbol{O}$ to extract OOD features $P_{\text{ood}} \in \mathbb{R}^{m\times n}$. Next, a loss $\mathcal{L}_{\text{dis}}$ is defined to maximize the distance between the virtual OOD features and ID features:

$$\mathcal{L}_{\text{dis}} = \text{KL}[q(\boldsymbol{O}), q(F)] + |p(P_{\text{ood}}) - p(P_{\text{in}})|, \tag{6}$$

where $q(\cdot)$ represents the probability distribution. By maximizing $\mathcal{L}_{\text{dis}}$, the distribution gap between $\boldsymbol{O}$ and F could be enlarged effectively, which promotes the synthesized features $\boldsymbol{O}$ to contain plentiful OOD information. Finally, to achieve the goal of distinguishing OOD objects from ID objects, P_{ood} and P_{in} are used to calculate an uncertainty loss [7], which regularizes the detector to produce a low OOD score for the ID object features, and a high OOD score for the synthesized OOD features:

$$\begin{aligned} \mathcal{L}_{\text{uncertainty}} = {} & \mathbb{E}_{u\backsim P_{\text{in}}}[-\log\frac{\exp^{-\text{E}(u)}}{1+\exp^{-\text{E}(u)}}] + \\ & \mathbb{E}_{v\backsim P_{\text{ood}}}[-\log\frac{1}{1+\exp^{-\text{E}(v)}}], \end{aligned} \tag{7}$$

where $\text{E}(\cdot)$ is the object-level energy score [7, 30]. During training, **the overall objective** is shown as follows:

$$\mathcal{L} = \mathcal{L}_{\text{in}} + \lambda \cdot (\mathcal{L}_{\text{cycle}} - \mathcal{L}_{\text{dis}}) + \tau \cdot \mathcal{L}_{\text{uncertainty}}, \tag{8}$$

where λ and τ are two hyper-parameters, which are set to 0.001 and 0.1 in the experiments.

3.4. Inference for OOD Object Detection

During inference, we only use the LoG operator to perform structure enhancement. And the operations of recurrent VAE and synthesizing virtual OOD features are not utilized in the inference stage. Besides, we only calculate the uncertainty loss for OOD object detection [7]. Specifically, for a predicted bounding box **b**, the processes of distinguishing OOD objects are shown as follows:

$$\mathcal{S} = \frac{\exp^{-\mathrm{E}(\mathbf{b})}}{1+\exp^{-\mathrm{E}(\mathbf{b})}}, \quad \mathcal{C}(\mathbf{b}) = \begin{cases} 0 & \text{if } \mathcal{S} < \gamma, \\ 1 & \text{if } \mathcal{S} \geq \gamma. \end{cases} \tag{9}$$

For the output of the classifier $\mathcal{C}(\cdot)$, we use the threshold mechanism [7] to distinguish the ID objects (the result is 1) from the OOD objects (the result is 0). The threshold γ is commonly set to 0.95 so that a high fraction of ID data is correctly classified. Finally, Algorithm 1 shows the training and evaluation details of our method.

Algorithm 1 SR-VAE for Unsupervised OOD-OD

```
Input: ID data $\{X, Y\}$, randomly initialized detector with parameter $\theta$, weight $\alpha$ for the $KL$-loss, weight $\lambda$ for the loss $\mathcal{L}_{\text{dis}}$, weight $\tau$ for the uncertainty loss $\mathcal{L}_{\text{uncertainty}}$.
Output: Object detector $\theta^*$, and OOD classifier $\mathcal{C}$.
while train do
    Sample images from the ID dataset $\{X, Y\}$.
    Calculate the structure-enhanced map $E$ and diverse augmented features $H$ using Eq. (2) and (3).
    Synthesize the virtual OOD map $\boldsymbol{O}$ using Eq. (5).
    Calculate the overall training objective $\mathcal{L}$ using Eq. (4), (6), (7), and (8).
    Update the parameters $\theta$ based on Eq. (8).
end
while eval do
    Calculate the OOD uncertainty score using Eq. (9).
    Perform thresholding comparison using Eq. (9).
end
```

4. Experiments

In the experiments, for unsupervised OOD-OD, we first evaluate our method on two different OOD object detection benchmarks [7]. Then, to further demonstrate the effectiveness of our method, we verify our method on the task of open-vocabulary detection (OVD) [33, 49] that aims to detect new classes defined by an unbounded vocabulary. Finally, we evaluate our method on the task of class-incremental object detection (IOD) [22], i.e., new classes are sequentially introduced into the object detector.

4.1. Experimental Setup

Implementation details. We utilize Faster R-CNN [36] with RoI-Alignment layer [12] as the basic detection model. ResNet-50 [13] is taken as the backbone. The weights pretrained on ImageNet [37] are used for initialization. Besides, in Eq. (3), $\Phi_\mu(\cdot)$ and $\Phi_\sigma(\cdot)$ separately consist of two fully-connected layers. The iteration number T is set to 3. In Eq. (5), W_μ and W_σ contain two convolutional layers, respectively. All the experiments are trained using the standard SGD optimizer with a learning rate of 0.02.

Datasets. For unsupervised OOD-OD, PASCAL VOC [8] and Berkeley DeepDrive (BDD-100k) [48] datasets are taken as the ID training data. Meanwhile, we adopt MS-COCO [29] and OpenImages [24] as the OOD datasets to evaluate the trained model. And the OOD datasets are manually examined the OOD images to ensure they do not contain ID category. For open-vocabulary detection, we follow the work [33] and conduct experiments on COCO [29]. Concretely, COCO-2017 dataset is used for training and validation. Meanwhile, 48 categories are selected as base classes and 17 are selected as new classes. For incremental object detection, we follow the standard evaluation protocol [22] and evaluate our method on PASCAL VOC [8]. We initially learn 10, 15, or 19 base classes, and then introduce 10, 5, or 1 new classes as the second task.

Metrics. To evaluate the performance of unsupervised OOD-OD, we report: (1) the false positive rate (FPR95) of OOD objects when the true positive rate of ID objects is at 95%; (2) the area under the receiver operating characteristic curve (AUROC); (3) mean average precision (mAP). For open-vocabulary detection and incremental object detection, we only report the mAP performance.

4.2. OOD-OD Performance Analysis

Table 1 shows the performance of unsupervised OOD-OD. We can see that though all methods own similar detection performance, the performance of OOD object detection differs significantly. This shows that existing detection methods are easily affected by OOD objects. Besides, for these two benchmarks, our method significantly outperforms baseline methods. Particularly, taking BDD [48] as the ID training data, based on FPR95 metric, our method separately outperforms VOS [7] by **12.04%** and **13.73%**. Meanwhile, based on AUROC metric, our method is 3.82% and 5.03% higher than VOS [7]. These results indicate that our method could indeed reduce the risk of misclassifying ID objects into OOD categories. Meanwhile, the proposed method could effectively synthesize virtual OOD features that deviate from the distribution of ID features, which improves the ability of distinguishing OOD objects.

In Fig. 4 and 5, we show some OOD detection examples. We can see that compared with the baseline method [7], our method accurately distinguishes OOD objects. Taking the first two columns of Fig. 4 as examples, the baseline method misclassifies OOD objects into ID categories. Our method correctly discriminates these objects as the OOD category.

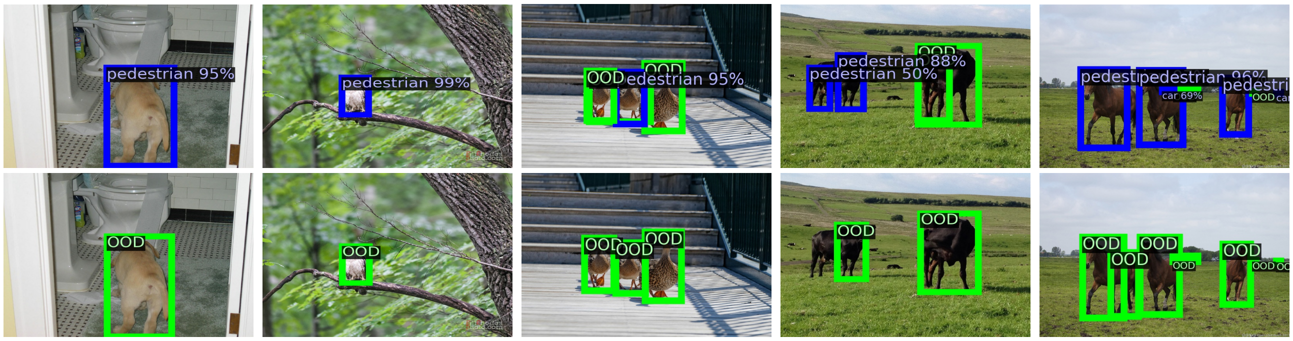


Figure 4. Detection results on the OOD images from MS-COCO. The first and second rows respectively indicate results based on VOS [7] and our method. The in-distribution dataset is BDD-100k. Blue boxes represent objects detected and classified as one of the ID categories. Green boxes indicate OOD objects. We can see that our method accurately determines OOD objects.

Method (VOC)	**FPR95 ↓**	**AUROC ↑**	**mAP (ID)↑**
	OOD: MS-COCO / OpenImages		
MSP [14]	70.99 / 73.13	83.45 / 81.91	48.7
ODIN [28]	59.82 / 63.14	82.20 / 82.59	48.7
Mahalanobis [26]	67.73 / 65.41	81.45 / 81.48	48.7
Gram matrices [38]	62.75 / 67.42	79.88 / 77.62	48.7
Energy score [30]	56.89 / 58.69	83.69 / 82.98	48.7
Generalized ODIN [16]	59.57 / 70.28	83.12 / 79.23	48.1
CSI [42]	59.91 / 57.41	81.83 / 82.95	48.1
GAN-synthesis [25]	60.93 / 59.97	83.67 / 82.67	48.5
VOS (Baseline) [7]	47.53 / 51.33	88.70 / 85.23	48.9
SR-VAE	**42.17 / 46.26**	**90.28 / 87.89**	**49.4**
Method (BDD)	**FPR95 ↓**	**AUROC ↑**	**mAP (ID)↑**
	OOD: MS-COCO / OpenImages		
MSP [14]	80.94 / 79.04	75.87 / 77.38	31.2
ODIN [28]	62.85 / 58.92	74.44 / 76.61	31.2
Mahalanobis [26]	55.74 / 47.69	85.71 / 88.05	31.2
Gram matrices [38]	60.93 / 77.55	74.93 / 59.38	31.2
Energy score [30]	60.06 / 54.97	77.48 / 79.60	31.2
Generalized ODIN [16]	57.27 / 50.17	85.22 / 87.18	**31.8**
CSI [42]	47.10 / 37.06	84.09 / 87.99	30.6
GAN-synthesis [25]	57.03 / 50.61	78.82 / 81.25	31.4
VOS (Baseline) [7]	44.27 / 35.54	86.87 / 88.52	31.3
SR-VAE	**32.23 / 21.81**	**90.69 / 93.55**	31.5

Table 1. The performance (%) of unsupervised OOD-OD. All methods are trained based on ID data and do not use any auxiliary data. ↑ denotes larger values are better and ↓ denotes smaller values are better. We can see that our method outperforms the comparison methods significantly.

These results further show that the proposed method could effectively synthesize OOD features and improve the discrimination ability of the object detector.

4.3. Performance Analysis of OVD and IOD

To demonstrate the superiorities, we further verify our method on OVD [49] and IOD [22]. Table 2 shows the OVD results on COCO. Here, we directly plug our method into the baseline method [33] and do not calculate the uncertainty loss (Eq. (7)). We can see that plugging our method improves OCA's performance by 3.5%. This shows that our method effectively synthesizes virtual features, which enhances the ability of distinguishing new classes.

Table 3 shows the IOD performance on PASCAL VOC. We still directly plug our method into the baseline method [22] and do not calculate the uncertainty loss. We can see that when the IOU threshold is set to 0.5 and 0.75, plugging our method effectively improves the performance of the baseline method. Particularly, when the IOU threshold is set to 0.75, for the three different settings, our method separately boosts iOD's performance by 2.3%, 2.9%, and 4.1%. These results show that our method could significantly improve the discrimination of the object detector.

Method	**AP$_{\text{novel}}$**	**AP**
WSDDN [3]	19.7	19.6
Cap2Det [47]	20.3	20.1
OVR-CNN [49]	22.8	39.9
RegionCLIP [50]	26.8	47.5
Detic [51]	27.8	45.0
OCA (Baseline) [33]	36.6	49.4
OCA+Ours	**40.1**	**49.5**

Table 2. OVD results (%) on COCO. 'OCA + Ours' indicates that we directly plug our method into OCA [33].

4.4. Ablation and Visualization Analysis

In this section, we utilize BDD-100k as the ID data for training and MS-COCO as the OOD data to perform an ablation analysis of our method.

Analysis of SR-VAE. In this paper, our method mainly contains the LoG operator used for enhancing object-related information, the VAE module used to generate diverse augmented features of the classification features, and the cycle-consistent conditional VAE module used for synthesizing virtual OOD features. Here, we make an ablation analysis of our method. Table 4 shows the ablation results. We can see that employing the proposed three modules could significantly improve the performance of OOD object detection. Particularly, compared with VOS [7], synthesizing virtual OOD features improves the performance by around 5.21% based on the FPR95 metric, which shows that the generated virtual OOD features could alleviate the impact of lacking unknown data for supervision and enhance the ability of distinguishing OOD objects. Meanwhile, employing the

10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	**mAP**
Faster ILOD (50) [32]	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.2
ORE (50) [19]	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77	67.7	64.6
OW-DETR (50) [10]	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
ROSETTA (50) [45]	74.2	76.2	64.9	54.4	57.4	76.1	84.4	68.8	52.4	67.0	62.9	63.3	79.8	72.8	78.1	40.1	62.3	61.2	72.4	66.8	66.8
iOD (50) [22]	76.0	74.6	67.5	55.9	57.6	75.1	85.4	77.0	43.7	70.8	60.1	66.4	76.0	72.6	74.6	39.7	64.0	60.2	68.5	60.5	66.3
iOD + Ours (50)	75.9	75.2	68.8	55.3	55.5	77.7	85.6	79.3	49.4	78.2	61.0	75.3	81.4	74.5	79.3	43.8	72.5	67.0	70.2	65.7	**69.6**
iOD (75) [22]	39.0	36.5	28.4	19.4	24.2	47.2	56.7	41.0	19.1	48.0	21.1	32.1	43.0	36.3	40.0	14.8	40.1	36.5	37.3	45.3	35.3
iOD + Ours (75)	43.6	41.0	31.3	24.9	29.8	55.4	60.8	44.1	22.4	46.7	29.5	32.3	35.6	38.3	35.7	15.1	46.9	34.6	37.9	46.7	**37.6**

15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	**mAP**
Faster ILOD (50) [32]	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
ORE (50) [19]	75.4	81.0	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
OW-DETR (50) [10]	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.4
ROSETTA (50) [45]	76.5	77.5	65.1	56.0	60.0	78.3	85.5	78.7	49.5	68.2	67.4	71.2	83.9	75.7	82.0	43.0	60.6	64.1	72.8	67.4	69.2
iOD (50) [22]	78.4	79.7	66.9	54.8	56.2	77.7	84.6	79.1	47.7	75.0	61.8	74.7	81.6	77.5	80.2	37.8	58.0	54.6	73.0	56.1	67.8
iOD + Ours (50)	78.3	80.3	70.5	51.6	60.2	79.4	85.9	76.2	52.5	79.4	65.2	81.8	83.7	76.1	77.9	41.1	62.8	63.8	72.6	67.9	**70.4**
iOD (75) [22]	40.7	40.9	28.7	19.1	23.8	61.6	56.1	38.8	23.6	47.5	18.7	40.1	40.2	41.5	39.8	9.1	40.6	32.4	41.9	47.6	36.6
iOD + Ours (75)	44.4	44.5	36.5	21.2	27.6	55.5	63.7	39.8	24.9	50.3	27.2	41.6	47.9	43.9	41.4	11.3	39.1	38.6	43.1	48.5	**39.5**

19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	**mAP**
Faster ILOD (50) [32]	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.6
ORE (50) [19]	67.3	76.8	60.0	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.9
OW-DETR (50) [10]	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	70.2
ROSETTA (50) [45]	75.3	77.9	65.3	56.2	55.3	79.6	84.6	72.9	49.2	73.7	68.3	71.0	78.9	77.7	80.7	44.0	69.6	68.5	76.1	68.3	69.6
iOD (50) [22]	78.2	77.5	69.4	55.0	56.0	78.4	84.2	79.2	46.6	79.0	63.2	78.5	82.7	79.1	79.9	44.1	73.2	66.3	76.4	57.6	70.2
iOD + Ours (50)	76.6	83.5	74.7	57.0	58.0	77.0	85.6	82.5	51.5	82.7	61.4	81.6	82.9	79.8	77.6	47.4	74.7	68.4	74.1	59.0	**71.8**
iOD (75) [22]	35.9	44.7	31.6	22.4	26.9	52.0	56.5	38.7	21.6	48.4	21.2	35.9	37.9	30.7	38.7	17.2	38.5	34.2	40.7	46.6	36.0
iOD + Ours (75)	36.4	45.1	36.1	18.0	28.9	53.2	62.2	38.5	25.3	55.1	27.4	46.8	45.9	42.9	40.3	20.9	50.8	37.0	44.4	47.1	**40.1**

Table 3. Performance (%) analysis of class-incremental object detection. 'iOD + Ours' indicates that our method is plugged into iOD [22]. Here, '50' and '75' separately represent that the mAP metric is calculated when the IOU threshold is set to 0.5 and 0.75.

C-VAE	LoG	R-VAE	**FPR95** ↓	**AUROC** ↑	**mAP**↑
✓			39.06%	87.83%	31.3%
✓	✓		37.92%	88.08%	31.2%
✓		✓	35.13%	89.22%	31.3%
✓	✓	✓	**32.23%**	**90.69%**	**31.5%**

Table 4. Ablation analysis of SR-VAE for unsupervised OOD-OD. 'C-VAE' is the proposed cycle-consistent conditional VAE module. 'R-VAE' indicates the VAE module that is used to generate diverse augmented features of the classification features.

LoG operator and R-VAE modules further boosts the performance significantly, which demonstrates that using these two modules is beneficial for enhancing object-related information in the extracted low-level features and improving the discrimination ability of the object detector.

Analysis of the LoG kernel size. To improve the performance of object localization, we employ the LoG operation (Eq. (1) and (2)) to perform structure enhancement, which is beneficial for strengthening object-related information. Here, we make an ablation analysis of the LoG kernel size. And we only change the kernel size. The other modules are kept unchanged. Table 5 shows the results. We can see that the performance is affected by the kernel size. Small kernel size does not sufficiently capture object structure information, which weakens the performance. For our method, the performance of using 9 × 9 kernel is the best.

LoG Kernel Size	**FPR95** ↓	**AUROC** ↑	**mAP**↑
3 × 3	34.96%	87.94%	30.9%
5 × 5	33.85%	89.18%	31.3%
7 × 7	33.08%	89.76%	31.2%
9 × 9	**32.23%**	**90.69%**	31.5%
11 × 11	32.93%	90.35%	**31.6%**

Table 5. Ablation analysis of the LoG kernel size.

Analysis of the iteration number. To reduce the risk of misclassifying ID objects into the OOD category, we design a VAE module (Eq. (3)) to generate multiple augmented features of the classification features, which is instrumental in improving the discrimination ability of the object classifier. Here, we make an ablation analysis of the iteration number T. And we only change the iteration number and keep other modules unchanged. Table 6 shows the ablation results. We can see that generating multiple augmented features could improve the detection performance effectively. This shows that enhancing the classification ability of the object classifier is meaningful. Besides, we observe that when the iteration number is larger than a certain threshold, the performance could not be boosted effectively. The reason may be that the object classifier is a linear classifier, which could not sufficiently exploit these augmented fea-

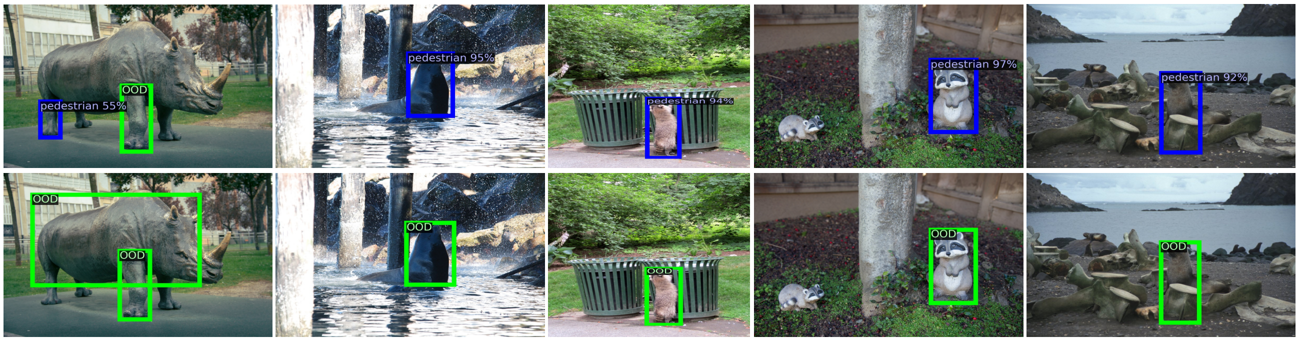
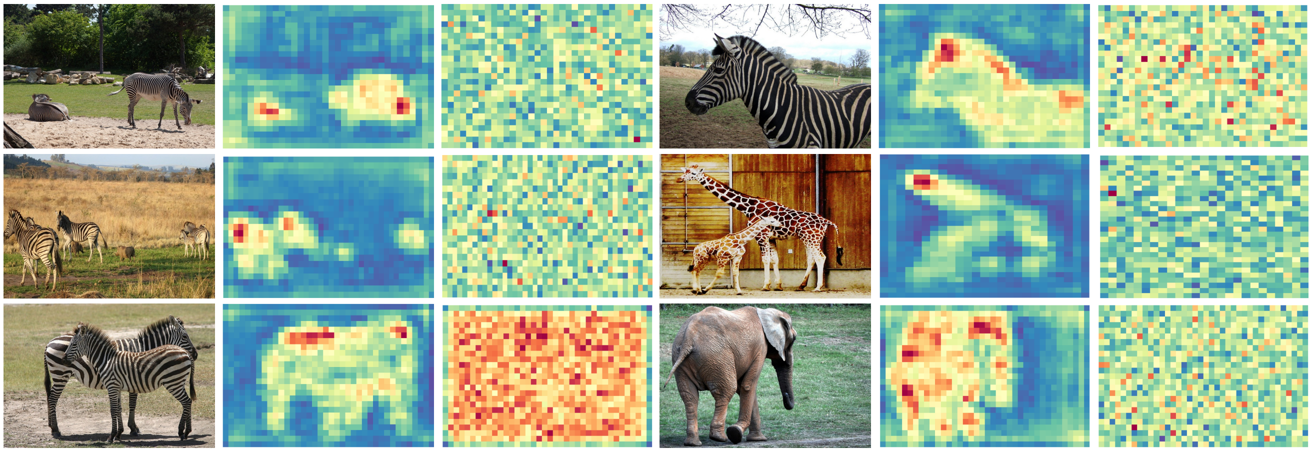


Figure 5. Detection results on OpenImages. The first and second rows respectively indicate results based on VOS [7] and our method.



(a) Input Image (b) Structure Map (c) Virtual OOD Map (d) Input Image (e) Structure Map (f) Virtual OOD Map

Figure 6. Visualization of the Structure-Enhanced map E (Eq. (2)) and Virtual OOD map $\boldsymbol{O}$ (Eq. (5)) based on the OOD data (MS-COCO). For each feature map, the channels corresponding to the maximum value are selected for visualization.

Iteration Number T	**FPR95** ↓	**AUROC** ↑	**mAP**↑
1	37.24%	88.12%	31.2%
3	**32.23%**	**90.69%**	**31.5%**
5	32.46%	90.23%	31.3%
7	32.93%	90.07%	31.2%

Table 6. Ablation analysis of the iteration number T.

tures. For our method, when the iteration number is set to 3, the performance of OOD detection is the best.

Analysis of cycle-consistent conditional VAE. In order to effectively synthesize virtual OOD features, we propose a cycle-consistency loss $\mathcal{L}_{\text{cycle}}$ (Eq. (8)) to reduce the impact of lacking paired samples. Here, we make an ablation analysis of $\mathcal{L}_{\text{cycle}}$. Based on the FPR95 metric, removing $\mathcal{L}_{\text{cycle}}$ increases the performance by around 3.8%. This shows using the cycle-consistency loss $\mathcal{L}_{\text{cycle}}$ is beneficial for synthesizing virtual OOD features and improving the ability of distinguishing OOD objects.

Visualization analysis. Fig. 6 shows some visualization examples. Particularly, we can see that the map E contains plentiful object-related information, which shows that using the LoG operation is indeed helpful for enhancing object-related information. Furthermore, we observe that the synthesized virtual OOD map $\boldsymbol{O}$ involves sufficient information that is different from object-related features, which indicates that leveraging the synthesized OOD features is instrumental in reducing the impact of lacking unknown data and improving the discrimination ability.

5. Conclusion

For unsupervised OOD-OD, we propose an approach of Structure-Enhanced Recurrent VAE. Particularly, to reduce the risk of misclassifying ID objects into the OOD category, an LoG operator and a dedicated recurrent VAE used to generate diverse augmented features are presented to strengthen the discrimination ability. Meanwhile, to alleviate the impact of lacking unknown data, we design a cycle-consistent conditional VAE to synthesize virtual OOD features, which boosts the performance of distinguishing OOD objects. Extensive experiments on three different tasks and visualization analyses demonstrate the superiorities of our method.

Acknowledgement. Our work was supported in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, the National Natural Science Foundation of China under Grant 62102293, 62132016, 62171343, and 62071361, and the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572, 2016. 2

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 6

[4] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ECCV*, 2020. 1

[6] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. *arXiv preprint arXiv:2203.03800*, 2022. 2

[7] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *ICLR*, 2022. 1, 2, 3, 4, 5, 6, 8

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 4

[10] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, pages 9235–9244, 2022. 7

[11] Ali Harakeh and Steven L Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. *arXiv preprint arXiv:2101.05036*, 2021. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 2, 3, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017. 2, 6

[15] 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. *ICLR*, 2017. 2, 3

[16] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, pages 10951–10960, 2020. 6

[17] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. *NeurIPS*, 31, 2018. 2

[18] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, pages 8710–8719, 2021. 2

[19] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 7

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[21] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 2

[22] Joseph Kj, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 5, 6, 7

[23] Hui Kong, Hatice Cinar Akakin, and Sanjay E Sarma. A generalized laplacian of gaussian filter for blob detection and its applications. *IEEE transactions on cybernetics*, 43(6):1719–1733, 2013. 2, 3

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International journal of computer vision*, 128(7):1956–1981, 2020. 2, 5

[25] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *ICLR*, 2018. 6

[26] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 31, 2018. 6

[27] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *CVPR*, pages 13218–13227, 2020. 2

[28] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *ICLR*, 2017. 6

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 33:21464–21475, 2020. 2, 4, 6

[31] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, volume 34, pages 5216–5223, 2020. 2

[32] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 140:109–115, 2020. 7

[33] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *NeurIPS*, 2022. 2, 5, 6

[34] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 1

[35] Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen. Anomaly detection requires better representations. *arXiv preprint arXiv:2210.10773*, 2022. 1

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 1, 2, 3, 5

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[38] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML*, pages 8491–8501. PMLR, 2020. 6

[39] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, pages 3400–3409, 2017. 2

[40] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *NeurIPS*, 28, 2015. 2, 3

[41] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, pages 14454–14463, 2021. 1

[42] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *NeurIPS*, 33:11839–11852, 2020. 6

[43] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *NeurIPS*, 33:19667–19679, 2020. 2

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 2

[45] Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. In *CVPR*, pages 9255–9264, 2022. 7

[46] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *ICCV*, pages 8301–8309, 2021. 2

[47] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *ICCV*, pages 9686–9695, 2019. 6

[48] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020. 5

[49] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 2, 5, 6

[50] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 6

[51] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 6

[52] Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. *CVPR*, 2022. 2