Can OOD Object Detectors Learn from Foundation Models?

Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi*

The University of Hong Kong
{liujh,wenxin,zhaosz,chenyx,xjqi}@eee.hku.hk

Abstract. Out-of-distribution (OOD) object detection is a challenging task due to the absence of open-set OOD data. Inspired by recent advancements in text-to-image generative models, such as Stable Diffusion, we study the potential of generative models trained on large-scale open-set data to synthesize OOD samples, thereby enhancing OOD object detection. We introduce SyncOOD, a simple data curation method that capitalizes on the capabilities of large foundation models to automatically extract meaningful OOD data from text-to-image generative models. This offers the model access to open-world knowledge encapsulated within off-the-shelf foundation models. The synthetic OOD samples are then employed to augment the training of a lightweight, plug-and-play OOD detector, thus effectively optimizing the in-distribution (ID)/OOD decision boundaries. Extensive experiments across multiple benchmarks demonstrate that SyncOOD significantly outperforms existing methods, establishing new state-of-the-art performance with minimal synthetic data usage. The project is available at https://github.com/CVMI-Lab/SyncOOD.

Keywords: OOD object detection · Synthetic data · Open-world data

1 Introduction

Modern object detectors, trained on closed-set data, have achieved remarkable success. However, they often incorrectly and confidently classify out-of-distribution (OOD) object categories as in-distribution (ID) categories in open-world applications [10], raising concerns about their reliability for deployment. To enhance the trustworthiness of object detection, researchers have studied the OOD object detection task, which aims to identify and flag unknown or novel objects as distinct from ID ones [3, 42].

The vulnerability of these models to OOD samples stems from their lack of awareness of the unknown open data distribution during training. Consequently, synthesizing OOD samples for model learning has emerged as a major research direction for this task [15, 59–61]. Most existing studies [15, 60, 61] concentrate on generating OOD objects in the latent space of an object detection model

* corresponding author

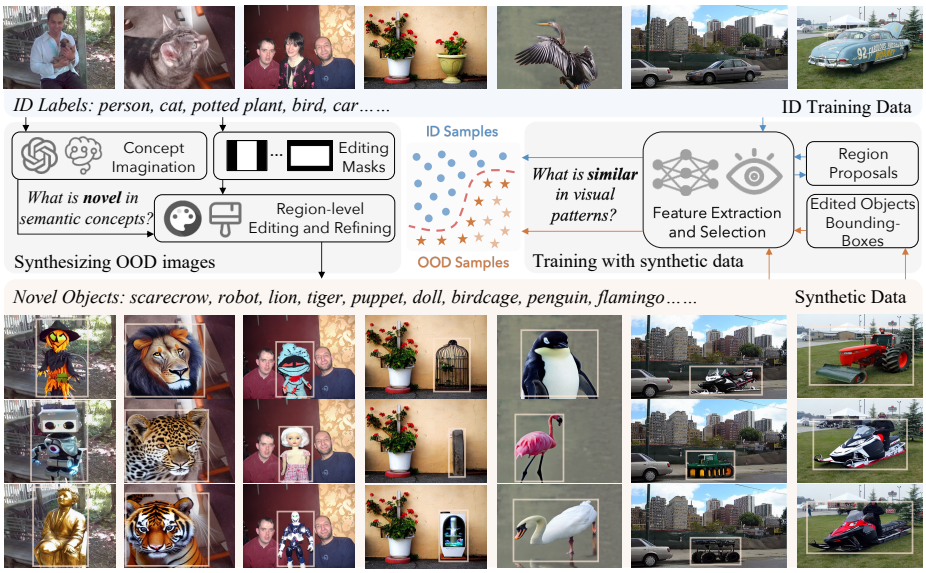

Fig. 1: Our pipeline replaces ID objects with semantic-novel yet visual-similar objects for scene-level OOD object synthesis. Middle left: The concepts are imagined by an LLM to ensure semantic separability and rationality, and reformed as text prompts for controllable in-painting using Stable Diffusion. Middle right: During training, only visually similar OOD objects are adopted based on instance-level feature similarity to the original object. A lightweight binary classifier is optimized for OOD detection, and other parts of the detector are kept unchanged.

trained on ID data. These synthesized samples are then used to optimize the decision boundary between ID and OOD data. Alternative approaches include directly synthesizing images by injecting adversarial noise [59] or identifying OOD instances from video data [14]. While these methods have yielded promising results, they remain limited to a closed-set setting, where the latent space for synthesizing outliers or the data is derived from a closed-set data distribution. Consequently, they may be biased towards the ID dataset, leading to suboptimal performance. Besides, effectively handling the unknown may appear unattainable when the unknown is never fully exploitable. Beyond that, *is it possible to learn them from massive open-world data knowledge condensed in foundation models?*

To this end, we investigate text-to-image generation models trained on a large amount of open-set data, which have demonstrated a superior ability to capture the distribution of data across a wide range of visual concepts, in order to synthesize novel data samples for enhancing OOD detection. Nonetheless, automatically extracting meaningful data from generative models for OOD object detection remains challenging due to the extensive vocabulary space to be explored, the complex scene-level synthesis problem, the need for object-level annotations, and potential distractions from contextual cues.

We introduce SyncOOD, an automatic data curation process that leverages foundation models as tools to harvest meaningful data from text-to-image gen-

eration models for OOD object detection (see Figure 1). The process is based on two key observations: 1) Hard OOD samples that are close to the ID data contribute more to learning a better OOD detector, and 2) Context may become a distracting cue for OOD object detection tasks, leading to bias towards contexts. With these observations in mind, the outlier synthesis process is formulated as box-conditioned image in-painting, and driven by Stable Diffusion [49] for high-quality controllable editing. The concepts to replace with are imagined by a Large Language Model (LLM) [1] with the aim of semantic novelty,[1] and the associated bounding box is further refined with SAM [30]. Automated by foundation models, this data collection pipeline requires minimal human labor, while producing high-quality OOD data.

In comparison to existing methods for OOD object detection, our core insight is to broaden the model's exposure to a more extensive range of open-set data and circumvent dataset biases by tapping into the open-world knowledge found in off-the-shelf foundation models. Utilizing generative models also provides us with control over the context of synthesized images and the data distribution. Our comprehensive experiments demonstrate superior performance, emphasizing the untapped potential of text-image-generation models in the context of OOD object detection. Our key contributions are summarized as follows:

- We investigate and unlock the potential of text-to-image generative models trained on large-scale open-set data for synthesizing OOD objects in object detection tasks.
- We introduce an automated data curation process for obtaining controllable, annotated scene-level synthetic OOD images for OOD object detection, which utilizes LLMs for novel concept discovery and visual foundation models for data annotation and filtering.
- We discover that maintaining ID/OOD image context consistency and obtaining more accurate OOD annotation bounding boxes are crucial for synthesized data to be effective in OOD object detection.
- Comprehensive experiments on multiple benchmarks demonstrate the effectiveness of our method, as we significantly outperform existing state-of-the-art approaches while using minimal synthetic data.

2 Related Work

OOD Object Detection For detecting OOD objects in scene-level images, unlike earlier works that constrain ID samples to a hypothetical distribution [12], it has become a recent trend to explicitly synthesize the outlier data, and incorporate them into training to adjust models' decision boundaries. However, due to the complexity of scene-level images, all previous works bypassed photo-realistic outlier synthesis in pixel space, and worked on generating outliers from the model's latent space [15, 60, 61], adversarial attack [59], or utilizing video data in the wild [14]. In the former, outliers can be sampled from the latent

[1] Concepts overlapping with the test data are removed to avoid information leakage.

space using a simple Gaussian prior [15], or more advanced generative models like VAE [61] or diffusion model [60]. Yet their upperbound are commonly limited by the quality of the latent space, and synthesized samples lack interpretability. Despite this, adversarial samples [59] lack semantic diversity, and auxiliary pseudo supervision from videos [14] introduces additional requirements to the setting. Unlike all above, our method applies LLM for OOD concept sampling, and Stable Diffusion for controllable image editing. This decouples the sampling and generation processes, elevates both parts to an unprecedented level, and achieves photo-realistic scene-level OOD image synthesis for the first time.

Open-world Object Detection The fact that real-world applications require object detectors the ability to tackle open categories is also considered in open-world object detection. The focus of this field includes generalization to domain shifts [57], incremental learning of novel classes [27, 39, 43], and zero-shot classification of open-vocabularies [17, 46]. Meanwhile, some works [18, 28] require to distinguish known objects well, be able to detect unknown objects, and finally be able to incrementally learn new objects. These works provide different perspectives on facing the challenges of real-world applications, which complement OOD object detection as a joint effort, but are out of the scope of this paper.

OOD Image Classification Earlier paradigms for OOD image classification either post hoc adjust models' confidence score at the testing phase, or apply regularization at models' training phase. The former line mainly focuses on the design of score functions, including confidence-based [2, 22, 35], energy-based [38, 56, 62], distance-based [34, 47, 50, 51], gradient-based [26], and approximating Bayesian [7, 10, 19, 40, 41] methods. The latter line of work includes regularizing models to produce lower confidence [23, 33], higher energy [29, 38], or directly shaping latent representations [12]. While outlier synthesis has shown to be effective by [15, 54], these are still generated in the latent space, and a parallel line study utilizing natural images [11, 23, 29] from the wild. Recently, photo-realistic outlier synthesis was first achieved by [13] with help from a text-conditioned diffusion model [49]. However, it is not readily applicable for object detection due to the complexity of scene-level images and the requirement for object-level annotations. In contrast, our work studies under the detection setting and requires outlier synthesis at scene level.

Foundation Models The evolution of large language models (LLMs) began with training on web-scale datasets [9, 45], leading to increasingly powerful foundation models [4, 6] capable of harnessing vast open-world data. Notable advancements include models [1] that interact with users and perform complex tasks like question answering, significantly broadening access to global knowledge. In image generation, diffusion models [24, 49] offer robust capabilities in synthesizing realistic content for applications such as image synthesis and inpainting. Additionally, segmentation foundation models like SAM [30] represent a leap forward in precise image segmentation, benefiting from extensive data training. Foundation models provide diverse data, which provides unlimited po-

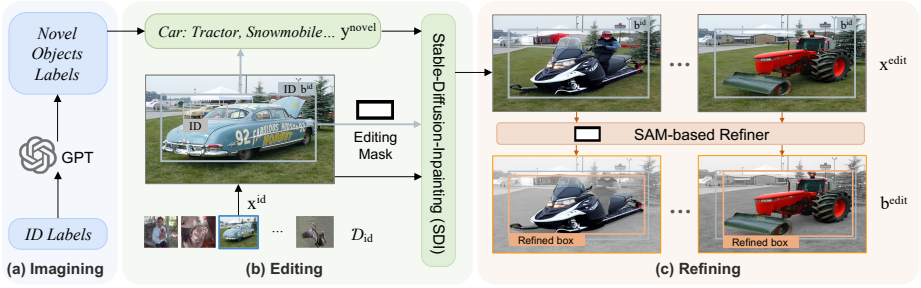

Fig. 2: Detailed illustration of our outlier synthesis pipeline. It comprises (a) Instructing an LLM to imagine semantic-novel concepts given ID objects, (b) Editing the selected regions to the expected concepts via prompt-conditioned image inpainting using Stable Diffusion, and (c) Refining the bounding boxes of edited objects using SAM.

tential for learning open-world knowledge from these models [55], where the effectiveness of data [21, 53, 63] is also crucial to the downstreaming tasks.

3 Method

Preliminary For OOD object detection, the training set contains only ID scene-level images $\mathbf{x}^{\text{id}}$ with ID objects, annotation bounding boxes $\mathbf{b}^{\text{id}}$, and semantic labels y, denoted as $\mathcal{D}_{\text{id}} = \left\{(\mathbf{x}^{\text{id}}, \mathbf{b}^{\text{id}}, \mathrm{y}^{\text{id}})\right\}$. The labels of ID objects always belong to a close set with K categories, denoted as $\mathrm{y}^{\text{id}} \in \mathcal{Y}_{\text{id}}$ and $\mathcal{Y}_{\text{id}} = \{\mathrm{y}_1^{\text{id}}, \mathrm{y}_2^{\text{id}}, ..., \mathrm{y}_K^{\text{id}}\}$. During inference, for each proposed object from an input scene-level image, it is required to identify whether its category belongs to $\mathcal{Y}_{\text{id}}$ or not.

Overview As illustrated in Fig. 1, our outlier synthesis pipeline consists of (1) synthesizing a set of effective photo-realistic scene-level OOD images $\mathbf{x}^{\text{edit}}$, denoted as $\mathcal{D}_{\text{edit}} = \left\{(\mathbf{x}^{\text{edit}}, \mathbf{b}^{\text{edit}})\right\}$, which contains novel objects and corresponding annotation boxes $\mathbf{b}^{\text{edit}}$ based on region-level editing from $\mathcal{D}_{\text{id}}$ in a fully automated, labor-free way; and (2) select and use the efficient synthetic data to provide pseudo-OOD supervisions for training OOD object detector together with the ID samples in the training set. Further design of the pipeline requires answering the following questions: (1) how to distill the open-set knowledge embedded in foundation models to scene-level OOD data and (2) how to utilize the synthesized data to regularize the decision boundary and facilitate OOD object detection. We discuss them accordingly in Sec. 3.1 and Sec. 3.2.

3.1 Synthesizing Semantic-novel Objects in Scene Images

Imagining Novel Concepts from ID objects As shown in Fig. 2(a), based on the ID labels $\mathcal{Y}_{\text{id}}$ in training set $\mathcal{D}_{\text{id}}$, we consider that novel concepts that are different from ID categories can be potential candidates for generating OOD

objects. The next is to discover novel concepts that offer hard OOD samples sharing high visual similarity with ID samples and being contextually compatible with the scene context for object detection. Rather than relying on human labor to investigate all potential candidates, we leverage the vast knowledge and reasoning capabilities of LLM, GPT-4 [1] to check the visual similarity and contextual compatibility. This allows us to associate ID objects and promote the conceptualization of possible novel objects to replace existing ID objects through the use of a prompt with in-context examples [4] as:

> Here is a list containing several objects $\mathcal{Y}_{\text{id}}$. Now, if I provide you an object name, you should return to me objects that are similar to the usage scenario and volume of the provided object but are not in the previous object list. For example, if I give you the word: person, you should respond and only respond: 'mannequin', 'sculpture', 'scarecrows', 'doll', 'puppet'.

With its robust logical foundation and rich knowledge, the LLM envisions a collection of novel objects for each ID object label, denoted as $\mathcal{Y}_{\text{novel}}$, while maintaining the semantic separability between imagined objects and ID objects. We empirically find one in-context example that is sufficient for us to discover novel concepts. For each ID label i, we discover M novel concepts using LLM $\mathbf{y}_i^{\text{novel}}$ of M concepts.

Editing Objects on Selected Regions With the discovered novel concepts $\mathcal{Y}_{\text{novel}} = \{\mathbf{y}_1^{\text{novel}}, \mathbf{y}_2^{\text{novel}}, ..., \mathbf{y}_K^{\text{novel}}\}$, the next step is to use them as prompts for the text-to-image generation model to generate an image. To generate a new image with novel concepts $y_j \in \mathbf{y}_i^{\text{novel}}$, we choose to replace existing ID objects in existing images with label y_i^{id} instead of finding new locations or generating one image from scratch. By doing so, we can ensure context compatibility and eliminate distractions from the scene context as it is preserved. As illustrated in Fig. 2 (b), we use Stable-Diffusion-Inpainting [49], denoted as SDI$(\cdot)$, to perform region-level editing on ID images. The ID object is denoted as $\mathbf{x}^{\text{id}}$, with its corresponding annotation box $\mathbf{b}^{\text{id}}$ serving as the editing mask, and the associated imagined novel concept $\mathbf{y}^{\text{novel}}$ are provided as inputs to the SDI, which is one of the most successful models for conditional image generation and editing. Thus, an edited image $\mathbf{x}^{\text{edit}}$ containing a novel object is obtained as:

$$\mathbf{x}^{\text{edit}} = \text{SDI}(\mathbf{x}^{\text{id}}, \mathbf{b}^{\text{id}}, \mathbf{y}^{\text{novel}}). \tag{1}$$

Refining Annotation Boxes of Novel Objects Due to the randomness in diffusion models, the attributes of edited objects, such as their quality, volume, and localization, may not match the original object box. To address this issue, as depicted in Fig. 2(c), we design an efficient and effective refiner based on SAM [30] to obtain refined accurate bounding boxes on novel objects. First, for an edited image $\mathbf{x}^{\text{edit}}$ with the editing mask $\mathbf{b}^{\text{id}}$, we use a padding area extended from $\mathbf{b}^{\text{id}}$ as the prompt and employ SAM to output the instance mask with

highest confidence $\mathbf{m}^{\text{SAM}}$ for the novel object in the area:

$$\mathbf{m}^{\text{SAM}} = \text{SAM}(\mathbf{x}^{\text{edit}}; \text{padding}(\mathbf{b}^{\text{id}}, e)), \tag{2}$$

where e represents the range of padding. Then, we convert obtained masks $\mathbf{m}^{\text{SAM}}$ to boxes $\mathbf{b}^{\text{SAM}}$, and calculate IoU between $\mathbf{b}^{\text{SAM}}$ and the corresponding $\mathbf{b}^{\text{id}}$ to filter out novel objects that vary highly in scale:

$$\left\{\mathbf{b}^{\text{edit}}\right\} = \left\{ \mathbf{b}^{\text{SAM}} \middle| \text{IoU}(\mathbf{b}^{\text{SAM}}, \mathbf{b}^{\text{id}}) > \gamma \right\}, \tag{3}$$

where γ denotes a threshold value on IoU. It ensures a high enough recall rate to rule out the instability of Stable Diffusion and SAM and uncontrollable localization of the edited objects. Thus we obtain the synthetic outlier data $\mathcal{D}_{\text{edit}}$ as illustrated in Fig. 1.

3.2 Mining Hard OOD Samples and Model Training

Mining Hard OOD Objects with High Visual Similarities for Training We consider the novel objects that are most likely to be confused with the corresponding ID objects by the object detector as the most effective ones. We thus aim to find synthetic OOD samples that are most easily confused as ID to participate in training based on pairwise similarity in the latent space of the pre-trained object detector. For each novel object with bounding box $\mathbf{b}^{\text{edit}}$ in the synthetic data $\mathcal{D}_{\text{edit}}$, we construct it with the corresponding original ID object with its bounding box as a pair: $\left\{(\mathbf{b}^{\text{edit}}, \mathbf{x}^{\text{edit}}), (\mathbf{b}^{\text{id}}, \mathbf{x}^{\text{id}})\right\}$. For an off-the-shelf object detector, denoted by $\mathcal{F}_{\text{det}}$, we extract latent features, $\mathbf{z}^{\text{edit}}$ and $\mathbf{z}^{\text{id}}$, for each pair:

$$\mathbf{z}^{\text{edit}}, \mathbf{z}^{\text{id}} = \mathcal{F}_{\text{det}}(\mathbf{b}^{\text{edit}}; \mathbf{x}^{\text{edit}}), \mathcal{F}_{\text{det}}(\mathbf{b}^{\text{id}}; \mathbf{x}^{\text{id}}). \tag{4}$$

The most effective novel objects are those with visual patterns that can be easily mistaken for their corresponding ID objects by an object detector. Therefore, we filter these novel objects based on their similarity to provide pseudo-OOD supervision:

$$\left\{\mathbf{z}^{\text{ood}}\right\} = \left\{ \mathbf{z}^{\text{edit}} \middle| \epsilon_{low} < \text{sim}(\mathbf{z}^{\text{edit}}, \mathbf{z}^{\text{id}}) < \epsilon_{up} \right\}, \tag{5}$$

where the similarities are computed between latent object features of edit-ID pairs. Here $\text{sim}(\cdot)$ denotes cosine similarity calculating and $\epsilon_{low}, \epsilon_{up}$ stand for the lower/upper similarity thresholds.

Optimizing ID/OOD Decision Boundary with Synthetic Samples Once we have obtained the ID and synthetic OOD objects, we employ a lightweight MLP, denoted as $\mathcal{F}_{\text{ood}}$, as the OOD detector optimized with a bi-classify loss:

$$\mathcal{L}_{\text{ood}} = \mathbb{E}_{\mathbf{z}\sim\mathbf{z}^{\text{id}}}\left[-\log \frac{1}{1+\exp^{-\mathcal{F}_{\text{ood}}(\mathbf{z})}}\right] + \mathbb{E}_{\mathbf{z}\sim\mathbf{z}^{\text{ood}}}\left[-\log \frac{\exp^{-\mathcal{F}_{\text{ood}}(\mathbf{z})}}{1+\exp^{-\mathcal{F}_{\text{ood}}(\mathbf{z})}}\right]. \tag{6}$$

The aforementioned design ensures both *semantic separability* and *pattern similarity* for the chosen synthetic samples. As a result, our proposed method elegantly optimizes the decision boundary using only a limited number of samples. It is further demonstrated and validated in the following experiments.

4 Experiments

Datasets Following OOD object detection setting [15], we use the **PASCAL-VOC** [16] and **Berkeley DeepDrive (BDD-100K)** [65] as the ID training datasets, which consist of 20 and 10 ID categories, respectively. Meanwhile, we evaluate the performance of our approach on two OOD datasets: **MS-COCO** [36] and **OpenImages** [32] respectively. Categories from the OOD datasets that overlapped with the ID datasets are removed to guarantee the absence of ID categories. We report ID categories, OOD categories, and texts used in driving image synthesis in the supplementary material.

Metrics We primarily focus on reporting the **FPR95** which represents the false positive rate of OOD samples when the true positive rate of ID samples is at 95% and lower is better. FPR95 is widely used to assess the OOD object detection performance [12, 15, 59–61]. Additionally, we present the Area Under the Receiver Operating Characteristic Curve (**AUROC**, higher is better) that is widely utilized to evaluate binary classification problems. Since we only train a plugin MLP on top of existing object detectors for OOD detection, ID performance, *e.g.*, mean Average Precision (mAP) is unchanged and thus omitted.

Implementation Details For our synthetic data, in order to maintain the experimental setting of OOD object detection and avoid leaking prior knowledge of foundation models, we **remove** all imagined novel objects that have the same or similar meaning as the ground truth OOD data categories. For model training, we follow the architectures of the baseline method [15, 59] to use a Faster R-CNN [48] as the base object detector with ResNet-50 [20] firstly. Then we trained a simple and lightweight 3-layer MLP. We follow [59] to extract multi-level features as training samples. ID samples are extracted from ID images and OOD samples are extracted from selected synthesized images with OOD bounding boxes. For training on the PASCAL-VOC dataset, we employ a learning rate of 1e-4, while for the BDD-100K dataset, we use a learning rate of 5e-5. Both training processes utilize a momentum of 0.9, a dropout rate of 0.5, and a batch size of 32. All training is conducted on GeForce RTX 3090 GPUs.

4.1 Main Results on OOD Object Detection

We evaluate the performance of the proposed method on different challenging benchmarks and obtain notable results (see Tab. 1). As the first work to introduce synthetic scene-level natural images as OOD samples, we incorporate our data-centric method to two off-the-shelf object detectors [15, 48], achieving new state-of-the-art performance in OOD object detection.

Compared with previous methods, we present comprehensive and substantial performance improvements on FPR95. The encouraging outcomes clearly show that our synthetic data offers superior OOD supervision and are well-suited for forming a precise decision boundary between ID and OOD samples as illustrated in Fig. 1, which significantly reduces the interference caused by contextual information when optimizing the decision boundary.

Table 1: Comparing on varied ID (PASCAL VOC [16], BDD-100K [65]) and OOD (MS-COCO [36], OpenImages [32]) datasets, our method significantly outperforms other methods on different metrics and achieves SOTA performance on OOD object detection. Our method is validated on two different existing object detectors, Faster R-CNN [48] and VOS [15] (denoted as Faster R-CNN + *Ours* and VOS + *Ours* respectively). (Top results are shown in **bold**.)

Method	ID:PASCAL-VOC				ID:BDD-100K			
	MS-COCO		OpenImages		MS-COCO		OpenImages	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP [22]	70.99	83.45	73.13	81.91	80.94	75.87	79.04	77.38
ODIN [35]	59.82	82.20	63.14	82.59	62.85	74.44	58.92	76.61
Mahalanobis [34]	96.46	59.25	96.27	57.42	57.66	84.92	60.16	86.88
Energy score [38]	56.89	83.69	58.69	82.98	60.06	77.48	54.79	79.60
Gram matrices [50]	62.75	79.88	67.42	77.62	60.93	74.93	77.55	59.38
Generalized ODIN [25]	58.57	83.12	70.28	79.23	57.27	85.22	50.17	87.18
CSI [52]	59.91	81.83	57.41	82.95	47.10	84.09	37.06	87.99
GAN-synthesis [33]	60.93	83.67	59.97	82.67	57.03	78.82	50.61	81.25
SIREN-KNN [12]	47.45	89.67	50.38	88.80	-	-	-	-
VOS [15]	47.53	88.70	51.33	85.23	44.27	86.87	35.54	88.52
SR-VAE [61]	42.17	90.28	46.26	87.89	32.23	90.69	21.81	93.55
DFDD [60]	41.34	**90.79**	44.52	88.65	30.71	90.74	22.67	92.48
SAFE [59]	47.40	80.30	20.06	92.28	32.56	88.96	16.04	94.64
Faster R-CNN + *Ours*	**36.44**	86.52	**13.34**	**95.37**	**22.67**	**95.44**	**12.96**	**96.26**
VOS + *Ours*	**34.97**	87.90	**11.25**	**96.96**	**23.09**	**94.32**	**14.12**	**96.41**

Table 2: Ablation on the number of our synthetic data in training. Taking PASCAL-VOC as the ID dataset, we perform seven groups of random sampling with different numbers in the synthetic dataset to extract features as OOD samples, evaluate and report the performance on the MS-COCO/OpenImages datasets.

#Sample	14k	12k	10k	8k	6k	4k	2k
FPR95↓	36.70/12.96	36.27/13.01	37.31/13.25	36.53/13.30	36.70/12.96	36.44/13.34	37.82/13.87
AUROC↑	86.65/95.54	86.68/95.55	86.64/95.51	86.69/95.52	86.75/95.56	86.52/95.44	86.03/95.18

Bridged by our synthetic data, foundation models' extensive knowledge and powerful logic about novel concepts are effectively injected into our model through novel concept imagining and region-level editing. Furthermore, powered by the similarity-based filter, our synthetic data proves to be highly effective. Compared with SAFE [59] which uses a similar framework, we only use around 25% (on PASCAL-VOC) and 20% (on BDD-100K) of auxiliary data to achieve a significant performance improvement. Further analysis is presented in Sec. 4.2.

4.2 Ablation Study

Number of Training Samples We conduct an extensive ablation study on the quantity of synthetic data utilized, as illustrated in Tab. 2. We employ seven sets of synthetic data with varying quantities as OOD samples, using PASCAL-VOC as the ID dataset. Features are extracted and the OOD detector is trained based on the same Faster R-CNN checkpoint for each set. It is noteworthy that

Table 3: We study the impact of associating varying numbers of imagined novel objects with each ID object. Taking PASCAL-VOC as the ID dataset, we report the performance on MS-COCO/OpenImages datasets.

#Sample	3	4	5	6	7	8
FPR95↓	36.96/13.58	37.13/13.15	37.31/12.82	36.87/13.25	37.91/13.53	37.13/13.15
AUROC↑	86.56/95.54	86.63/95.51	86.46/95.47	86.51/95.37	86.43/95.35	86.44/95.56

Table 4: We randomly sample the same numbers of OOD features as the main experiment instead of using the feature filter (denoted as w/o filter), and evaluate on multiple datasets. The obtained results demonstrate the effectiveness of the proposed data filter.

Method	ID:PASCAL-VOC				ID:BDD-100K			
	MS-COCO		OpenImages		MS-COCO		OpenImages	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
w/o filter	39.29	85.46	13.68	95.22	25.45	93.17	15.32	95.83
Ours	**36.44**	**86.52**	**13.34**	**95.37**	**22.67**	**95.44**	**12.96**	**96.26**

as the number of samples decreases from 14k to 2k, the performance does not deteriorate but rather maintains stable and superior results. This highlights our method's data efficiency (SAFE used 16k samples). Combined with our feature similarity-based filtering strategy as in Sec. 3.2, a small number of high-quality OOD samples with *visual similarity* directly promotes the optimization of precise decision boundaries to achieve stable improvements.

Meanwhile, we use the same detector checkpoint to assess the parallel baseline, SAFE [59], and get the performance of **50.86/23.60** on **FPR95** and **78.15/91.42** on **AUROC**. SAFE augments about 16k images and extracts more than 100k instance-level features as OOD samples. In contrast, our approach utilizes fewer synthetic images and extracts only one instance-level feature from the edited novel object in each synthetic image as an OOD sample, resulting in significantly superior performance compared to SAFE [59].

Number of Concepts to Imagine We employ in-context learning to guide LLM in associating new objects for driving image editing. For each ID obejct, the LLM connects a steady stream of novel objects. We further explore the impact of the number of corresponding novel objects for each ID object on data performance. We randomly sample different numbers of novel objects from LLM's responses for each ID object. As shown in Tab. 3, the performance remains stable despite changes in the number of concepts, further highlighting the stability and robustness of our synthetic data.

SAM-based Refiner As mentioned in Sec. 3.1, we propose to utilize SAM-based refiner to correct the bounding boxes of novel objects to obtain higher-quality instance-level OOD features. Therefore, we comparatively remove the proposed refiner and directly used the corresponding editing masks as bound-

Fig. 3: We show cases on six intervals of feature similarity (consistent with Eq. (5), indicated at the bottom of the figure). The first line contains the corresponding initial images, the second line contains the synthetic images with the corresponding boxes of the novel objects (yellow boxes), and the third line contains the difference heat maps of the latent feature maps extracted from the above image pairs (superimposed on the corresponding synthetic images, denoted as Diff-map).

ing boxes to extract OOD features for training. Taking PASCAL-VOC as the ID dataset, after removing the refiner, we obtain **39.55/13.72** of **FPR95** and **85.94/95.37** of **AUROC** on MS-COCO/OpenImages datasets, which is better than previous methods but worse than the results (Faster R-CNN + *Ours* in Tab. 1) when using the refiner. This proves that OOD supervision signals contained in the synthetic data are already extracted and achieve good results under the localization of the fuzzy boxes, but more precise boxes mean higher quality features. More demos and analyses of the SAM-based refiner are shown in the supplementary material.

Similarity-based Filter The filter is designed to incorporate the most useful data into training, and avoid unnecessary noise. The design is reflected in two aspects: on one hand, the outlier object should process similar visual patterns to the original object, thus being confusing and can facilitate learning; on the other hand, over-high similarity may indicate failures of the editing process (*e.g.*, when the concept is not an object). These considerations are applied as thresholding on pairwise cosine similarity between object features, as in Eq. (5). As shown in Tab. 4, this filter brings a notable improvement across benchmarks.

Table 5: Comparing varied images as OOD samples for training, we first show some synthetic object-centric images generated by Stable Diffusion (left side). Then with PASCAL-VOC as ID dataset, we report the results obtained by using synthetic object-centric images (denoted as object-centric images) and scene-level images with novel objects but without bounding boxes (denoted as scene-level w/o boxes) as OOD samples to participate in training.

Object-centric Images	Data	MS-COCO		OpenImages	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
	object-centric images	51.99	81.48	20.70	93.85
	scene-level w/o boxes	48.01	82.38	18.61	93.44
	Ours	**36.44**	**86.52**	**13.34**	**95.37**

To provide more insights on the choices of filtering thresholds, we display some cases in different intervals of feature similarity in Fig. 3. We show the synthetic images, the corresponding initial images, and the difference between feature maps (denoted as Diff-map), respectively. The Diff-maps prove that the edited area is sensitively attended to by the model. But for images with extremely high similarity (> 0.9), they always contain some editing failures and blurs. As illustrated on the top of the first column in Fig. 3, it is not intuitively evident what the *ship* has been edited into (the target object is a *raft*). Besides, as the similarity upperbound decreases, we progressively obtain more realistic and reasonable images. But note that when the similarity is excessively low, as seen in the last column of Fig. 3, the objects are edited into the corresponding text or an unnatural object, leading to image distortion. This strongly supports the idea that the quality and usability of edited images are closely connected to visual similarity. More cases and analyses are presented in the supplementary material.

4.3 Discussions on Outlier Synthesis

Scene-level Editing Matters Through regional-level editing, we replace the ID object with a novel object with a bounding box and ensure consistent context information. However, some simpler methods based on foundation models also achieve the acquisition and use of OOD data. For example, Dream-OOD [13] uses well-trained text-conditional space and diffusion model to synthesize realistic object-centric data for promoting OOD image classification. Similarly, keeping other settings unchanged, we use our novel concepts to drive Stable-Diffusion instead of Stable-Diffusion-Inpainting [49] to synthesize novel images, which are also processed and filtered by our proposed refiner and filter (some synthesized images are shown in Tab. 5), thereby participating in the training as OOD supervision. However, as shown in Tab. 5 (object-centric images), the synthetic novel object-centric images do not aid in training and result in poor performance, even though they possess high visual quality. This clearly validates our decision to edit scene-level images rather than composing new ones, and highlights the significance of maintaining contextual consistency.

Additionally, we examine the possibility of using the edited scene-level image as a whole (ignoring the boxes) as OOD samples in the training process. The

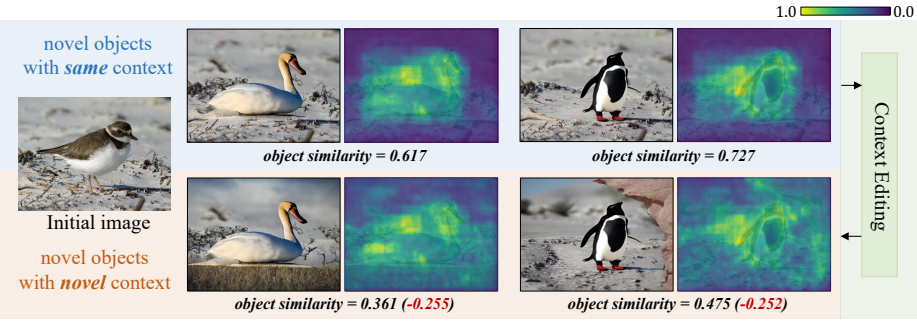

Fig. 4: We edit the context of the synthetic data (in blue box) so that the images contain novel objects and novel context (in orange box). Then we calculate the similarity between the instance-level feature of the corresponding objects from all synthetic images and the instance-level feature in the initial image (left, the bird). The similarities and the corresponding difference maps are shown in the figure.

results, as depicted in Tab. 5 (scene-level w/o boxes), are significantly inferior compared to our method's performance. This demonstrates that controllable bounding boxes are indispensable in this task.

Context Consistency Matters Given the importance of scene-level synthesis as discussed in the previous paragraph, we study the factors that make scene-level editing indispensable, and find context consistency to be a crucial one. Besides calculating the similarity between ID/OOD object pairs before/after box-conditioned editing, we also try further editing parts of the background of the already edited images, and also calculate its object similarity with the initial object. As shown in Fig. 4, even small editing on parts of the background (out of the object boxes) can make foreground objects 'look' notably different as perceived by the detector. This highlights the importance of keeping the context unchanged when synthesizing outlier samples, in that if the context is changed, the model easily identifies the object as OOD and cannot break the context bias.

5 Discussion

What Type of OOD Data Matters? We are the first to explore how to edit scene-level images to include novel categories, which contain annotation boxes and ensure context consistency, facilitating OOD object detection. The achieved state-of-the-art performance (Tab. 1) benefits from the optimization of decision boundaries driven by high-quality OOD features. Our exploration demonstrates that *annotation boxes* and *context consistency* are particularly important for synthesizing high-quality OOD instances. On the one hand, high-quality annotation boxes provide the possibility to extract high-quality instance-level features from scene-level images, while unrefined boxes (Sec. 4.2) or discarded boxes (Tab. 5) will have a negative impact on performance. On the other hand, context consistency ensures the most effective OOD features are not interfered by different contexts and selected for utilizing (Fig. 4).

How to Synthesize Suitable OOD Data? We are the first to build *an automatic, transparent, and low-cost pipeline* (Sec. 3.1) for synthesizing scene-level images containing novel objects with annotation boxes and context consistency. It benefits object detectors' robustness and reliability to unseen data and sets up clear state-of-the-art on multiple OOD object detection benchmarks. Specifically, we organically combine and cleverly use different foundation models [1, 30, 49] (Sec. 3.1) to distill open-world knowledge and inpaint the existing scenes for simulating real OOD scenarios. In addition, our design *takes into account the instability of the current foundation models* and can release better potential performance in the future development of foundation models.

How to Select Suitable OOD Data? We are the first to *explicitly decouple OOD data synthesis and selection*. On the one hand, we ensure the separability of the synthetic objects in semantic concepts through open-world knowledge provided by LLMs (Sec. 3.1). On the other hand, we ensure the similarity of ID and OOD objects in visual patterns (Sec. 3.1), thereby optimizing the precise decision boundary. This line of thinking has the potential to facilitate more open-world solutions.

Broader Impacts Beyond showcasing engineering success via effectively combining specific foundation models, our work uncovers the untapped potential of the text-to-image generative models and visual foundation models in pushing forward the OOD object detection task to effectively leverage the off-the-shelf open-world data knowledge [5,31,66]. More importantly, our work establishes a bridge between OOD object detection and the latest advancements in deep learning, enabling it to benefit from ongoing developments, go beyond isolated academic practice, and resolve practical challenges in open-world applications. Meanwhile, automating novel data generation and curation will inspire more tasks in more modalities, such as in visual-language [44, 58] and 3D vision [8, 37, 64].

6 Conclusion

In this paper, we investigate improving OOD object detection by distilling open-world data knowledge from text-to-image generative models. We develop an automatic and cost-effective data curation pipeline, SyncOOD, that leverages foundation models as tools to obtain meaningful open-set data from generative models. Through extensive studies, we discover that object boxes and context consistency of the generated data contribute to the improvement of OOD object detection performance. Our comprehensive experiments demonstrate that SyncOOD not only advances the state-of-the-art in OOD object detection but also emphasizes the untapped potential of utilizing large-scale generative models for enhancing the robustness of machine learning systems in open-world settings. As an initial exploration in leveraging foundation models for OOD object detection, we hope our promising results encourage further research in advancing this area in the future.

Acknowledgments

This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), Theme-based Research (Grant No. T45-701/22-R) and RGC Matching Fund Scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust. We would like to thank Chirui Chang, Xiaoyang Lyu, Haoru Tan, and Xiuzhe Wu for their insightful discussions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1563–1572 (2016)
3. Bishop, C.M.: Novelty detection and neural network validation. IEE Proceedings-Vision, Image and Signal processing **141**(4), 217–222 (1994)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
5. Chang, C., Liu, Z., Lyu, X., Qi, X.: What matters in detecting ai-generated videos like sora? arXiv preprint arXiv:2406.19568 (2024)
6. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
7. Deepshikha, K., Yelleni, S.H., Srijith, P., Mohan, C.K.: Monte Carlo dropblock for modelling uncertainty in object detection. arXiv preprint arXiv:2108.03614 (2021)
8. Deng, W., Ding, R., Yang, J., Liu, J., Li, Y., Qi, X., Ngai, E.: Can 3d vision-language models truly understand natural language? arXiv preprint arXiv:2403.14760 (2024)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dhamija, A., Gunther, M., Ventura, J., Boult, T.: The overlooked elephant of object detection: Open set. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1021–1030 (2020)
11. Du, X., Fang, Z., Diakonikolas, I., Li, Y.: How does unlabeled data provably help out-of-distribution detection? arXiv preprint arXiv:2402.03502 (2024)
12. Du, X., Gozum, G., Ming, Y., Li, Y.: SIREN: Shaping representations for detecting out-of-distribution objects. Advances in Neural Information Processing Systems **35**, 20434–20449 (2022)
13. Du, X., Sun, Y., Zhu, J., Li, Y.: Dream the impossible: Outlier imagination with diffusion models. Advances in Neural Information Processing Systems **36** (2024)

14. Du, X., Wang, X., Gozum, G., Li, Y.: Unknown-aware object detection: Learning what you don't know from videos in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13678–13688 (2022)
15. Du, X., Wang, Z., Cai, M., Li, Y.: VOS: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)
16. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. International journal of computer vision **88**, 303–338 (2010)
17. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022)
18. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9235–9244 (2022)
19. Hall, D., Dayoub, F., Skinner, J., Zhang, H., Miller, D., Corke, P., Carneiro, G., Angelova, A., Sünderhauf, N.: Probabilistic object detection: Definition and evaluation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1031–1040 (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
21. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574 (2022)
22. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
23. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10951–10960 (2020)
26. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Advances in Neural Information Processing Systems **34**, 677–689 (2021)
27. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
28. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5830–5840 (2021)
29. Katz-Samuels, J., Nakhleh, J.B., Nowak, R., Li, Y.: Training ood detectors in their natural habitats. In: International Conference on Machine Learning. pp. 10848–10865. PMLR (2022)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

31. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 813–822 (2021)
32. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
33. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations (2018)
34. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems **31** (2018)
35. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations, ICLR 2018 (2018)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
37. Liu, J., Chang, C., Liu, J., Wu, X., Ma, L., Qi, X.: Mars3d: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9372–9381 (2023)
38. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems **33**, 21464–21475 (2020)
39. Liu, X., Yang, H., Ravichandran, A., Bhotika, R., Soatto, S.: Continual universal object detection. arXiv preprint arXiv:2002.05347 (2020)
40. Miller, D., Dayoub, F., Milford, M., Sünderhauf, N.: Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In: International Conference on Robotics and Automation (ICRA). pp. 2348–2354 (2019)
41. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 3243–3249 (2018)
42. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 427–436 (2015)
43. Perez-Rua, J.M., Zhu, X., Hospedales, T.M., Xiang, T.: Incremental few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13846–13855 (2020)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
45. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
46. Rahman, S., Khan, S.H., Porikli, F.: Zero-shot object detection: Joint recognition and localization of novel concepts. International Journal of Computer Vision **128**, 2979–2999 (2020)

47. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A simple fix to Mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022 (2021)
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Sastry, C.S., Oore, S.: Detecting out-of-distribution examples with gram matrices. In: International Conference on Machine Learning. pp. 8491–8501. PMLR (2020)
51. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning. pp. 20827–20840. PMLR (2022)
52. Tack, J., Mo, S., Jeong, J., Shin, J.: CSI: Novelty detection via contrastive learning on distributionally shifted instances. Advances in Neural Information Processing Systems **33**, 11839–11852 (2020)
53. Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F., Qi, X.: Data pruning via moving-one-sample-out. Advances in Neural Information Processing Systems (2023)
54. Tao, L., Du, X., Zhu, X., Li, Y.: Non-parametric outlier synthesis. arXiv preprint arXiv:2303.02966 (2023)
55. Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., Isola, P.: Learning vision from models rivals learning vision from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15887–15898 (2024)
56. Wang, H., Liu, W., Bocchieri, A., Li, Y.: Can multi-label classification networks know what they don't know? Advances in Neural Information Processing Systems **34**, 29074–29087 (2021)
57. Wang, X., Huang, T.E., Liu, B., Yu, F., Wang, X., Gonzalez, J.E., Darrell, T.: Robust object detection via instance-level temporal cycle confusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9143–9152 (2021)
58. Wen, X., Zhao, B., Chen, Y., Pang, J., Qi, X.: Generalization beyond data imbalance: A controlled study on clip for transferable insights. arXiv preprint arXiv:2405.21070 (2024)
59. Wilson, S., Fischer, T., Dayoub, F., Miller, D., Sünderhauf, N.: SAFE: Sensitivity-aware features for out-of-distribution object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23565–23576 (2023)
60. Wu, A., Chen, D., Deng, C.: Deep feature deblurring diffusion for detecting out-of-distribution objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13381–13391 (2023)
61. Wu, A., Deng, C.: Discriminating known from unknown objects via structure-enhanced recurrent variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23956–23965 (2023)
62. Wu, Q., Chen, Y., Yang, C., Yan, J.: Energy-based out-of-distribution detection for graph neural networks. arXiv preprint arXiv:2302.02914 (2023)
63. Wu, S., Tan, H., Tian, Z., Chen, Y., Qi, X., Jia, J.: Saco loss: Sample-wise affinity consistency for vision-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27358–27369 (2024)

64. Wu, X., Dai, P., Deng, W., Chen, H., Wu, Y., Cao, Y.P., Shan, Y., Qi, X.: Cl-nerf: continual learning of neural radiance fields for evolving scene representation. Advances in Neural Information Processing Systems **36** (2024)
65. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2020)
66. Zheng, H., Wang, Q., Fang, Z., Xia, X., Liu, F., Liu, T., Han, B.: Out-of-distribution detection learning with unreliable out-of-distribution sources. Advances in Neural Information Processing Systems **36**, 72110–72123 (2023)