Improving Out-of-Distribution Detection Performance using Synthetic Outlier Exposure Generated by Visual Foundation Models

Gitaek Kwon*[1]
gitaek.kwon@vuno.co
Jaeyoung Kim*[2]
jaydee.kim@kakaohealthcare.com
Hongjun Choi[1]
hongjun.choi@vuno.co
Byungmoo Yoon[3]
byungmoo.yoon@gachon.ac.kr
Sungchul Choi[4]
sc82.choi@pknu.ac.kr
Kyu-Hwan Jung[5]
khwanjung@skku.edu

[1] VUNO Inc.,
Korea
[2] KakaoHealthcare Corp.,
Korea
[3] Gachon University,
Korea
[4] Pukyong National University,
Korea
[5] Samsung Advanced Institute for Health Sciences and Technology,
Sungkyunkwan University,
Korea

Abstract

Real-world deep learning applications often encounter out-of-distribution (OOD) samples that do not belong to the label spaces of the training dataset. Therefore, neural networks should detect OOD samples and refrain from making predictions on the detected ones to help users be less confused about models' decisions. A rejection network that has learned representations of OOD can be used to detect distribution shifts, but most existing methods require an additional data collection procedure to train the rejection network. In this paper, we propose the **S**ynthetic **H**armless outlier **I**mages generator **F**rom **T**raining samples (**SHIFT**), a realistic OOD generator that converts a training image into synthetic OOD samples by using vision foundation models in a zero-shot manner. Specifically, to construct the surrogate OOD image, the SHIFT uses CLIP to erase the regions of the in-distribution (ID) object, and the latent diffusion model replaces the key regions with realistic features considering the marginal background. Therefore, our method can eliminate the need to collect external outlier samples to train a rejection network. We demonstrate the competitiveness of the proposed method on several benchmarks (i.e., CIFAR-10/100 and STL-10), and the code is publicly available at https://github.com/Anears/SHIFT.

*These authors contributed equally.

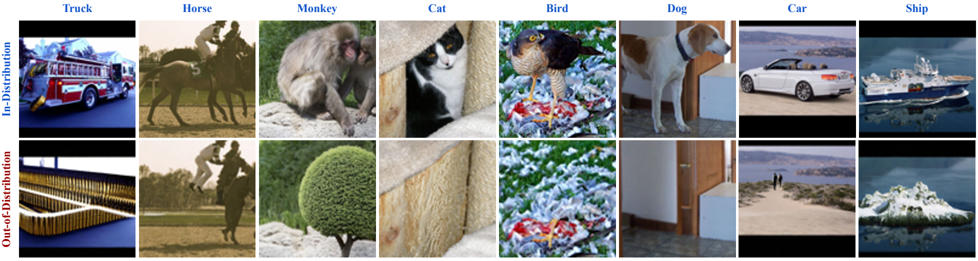

Figure 1: Examples of our synthetic OOD samples constructed from STL-10 (ID). Each row represents ID images and generated OOD images with removed ID objects, respectively.

1 Introduction

In a wide range of safety-critical applications, such as object recognition in autonomous driving [7] and computer-aided diagnosis systems [18, 33, 34], deep neural networks (DNNs) have shown remarkable success under the closed-world assumption. However, in many real-world deployments of neural networks, the test sample could potentially be sampled from arbitrary input space, and previous studies have shown that DNNs often produce unreliable predictions for out-of-distribution (OOD) samples that are outside of the training distribution [13, 30]. Since these erroneous predictions can confuse users when interpreting the model's decisions, recent DNNs require the ability to detect OOD samples and subsequently refrain from making predictions for such samples.

To address this problem, a widely adopted approach is to manipulate the model's outputs by leveraging scoring functions derived from a pre-trained classifier [11, 22, 25, 35, 36]. These scoring functions assign higher scores to the in-distribution (ID) examples, whereas lower scores to the OOD examples. Another popular approach for detecting OOD inputs is training a rejection network using auxiliary OOD datasets [3, 12, 27, 42]. Rejection networks exposed to diverse OOD representations excel at detecting OOD samples, but training the rejection network requires an additional burden for collecting external OOD datasets (sometimes infeasible to obtain in practice), and the optimal choice of external OOD datasets remains an open question [6].

Instead, several studies suggest synthetic OOD sample generation techniques that generate virtual OOD samples using a generative adversarial network (GAN) [21, 38] and adversarial training [10]. Although they show improved OOD detection performances on CIFAR-10/100 datasets, a rejection network trained on synthetic high-resolution data still perform poorly than when using real OOD examples due to distributional discrepancy between synthetic- and real-OOD samples [19].

In this paper, we propose a simple yet effective method, called **SHIFT**, to generate a *realistic* OOD sample by converting a training image into an OOD sample (see Figure 1). Our proposed procedure is based on contrastive language-image pre-training (CLIP) [31] and latent diffusion model (LDM) [32], inspired by the extraordinary capabilities of foundation models [1, 14, 26]; (1) we mask the ID object in the training image using the CLIP-based segmentation model [26], then (2) inpaint the masked region with its background by leveraging the LDM.

After constructing the OOD dataset, a rejection network operating on a pre-trained classifier's latent space can then be trained on the generated OOD and ID datasets in order to detect OOD data in a test time. Our main contributions are as follows:

- The OOD dataset is constructed offline while being adaptable to any ID training data for image classification tasks. Compared to the previous method using outlier exposure [12], our proposed method eliminates the need to collect an external OOD dataset.

- The rejection network trained with constructed OOD dataset shows state-of-the-art performances in the OOD detection task by improving averaged area under the receiver operating characteristics (AUROC) +0.9% (CIFAR-10), +2.5% (CIFAR-100), and +5.9% (STL-10), while preserving the classification accuracy on the ID dataset.

- To the best of our knowledge, this is the first study to propose how to generate OOD examples using visual foundation models in a zero-shot manner. Through comprehensive experiments, we confirm that the proposed method does indeed generate realistic OOD examples from a training image, and demonstrate its effectiveness in training a rejection network.

2 Related Work

Post-hoc methods are those that can detect OOD samples without re-training a pre-trained classifier. Hendrycks and Gimpel [11] present a simple baseline that utilizes the maximum softmax probability (MSP) as a scoring function. Subsequently, improved algorithms have been proposed; ODIN [23] shows that the addition of controlled perturbations to test inputs and temperature scaling can separate confidence scores between ID and OOD samples. Assuming the feature space of a pre-trained classifier as class-conditional Gaussian variables, Lee et al. [22] propose the Mahalanobis distance-based rejection rule. Liu et al. [25] use the energy score (Energy) as an OOD scoring rule to align with the probability density of the logits for a pre-trained network. DICE [35] is a sparsification-based OOD detection technique that ranks weights by contribution and then uses the most significant weights to reduce noisy signals in OOD samples. However, many post-hoc algorithms rely on tuning with a small subset of OOD data. It is not only hard to define a-priori, but also hyperparameters tuned using the subset of OOD data can lead to biased results [15].

Another line of approaches explores the rejection network using an additional outlier dataset. Outlier exposure (OE) [12] uses auxiliary datasets completely disjoint from ID classes to teach the model a representation for ID/OOD distinctions. However, in real-world applications, OE has a limitation in that collecting all possible OOD samples is not feasible. Instead, several studies propose an efficient method for generating a synthetic outlier sample; Lee et al. [21] use GAN to generate synthetic OOD samples that are close to training distribution but also simultaneously have high entropy in terms of classifier output over these samples. Their key finding is that synthetic OOD samples are most useful when they lie near ID examples on feature spaces of the pre-trained network. CEDA and ACET [9] both use random noise and pixel shuffling of ID samples, the latter including ad additional adversarial enhancement procedure. Kim et al. [19] suggest a novel approach called KIRBY which generates surrogate OOD data from training samples, they show the state-of-the-art OOD detection performance on the vision benchmarks.

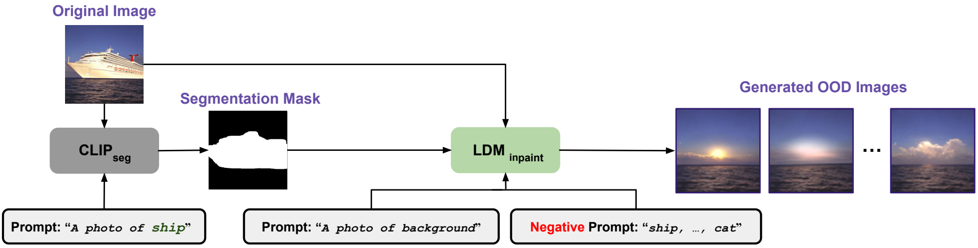
Original Image
Segmentation Mask
Generated OOD Images
CLIPseg
LDM inpaint
...
Prompt: "A photo of ship"
Prompt: "A photo of background"
Negative Prompt: "ship, …, cat"

Figure 2: Overall procedure of SHIFT.

3 Preliminary

Out-of-Distribution Detection. For image classification tasks, let $\mathcal{X} \subset \mathbb{R}^d$ be the input space and let $\mathcal{Y}$ be the label space. Denote by $\mathcal{D}_{\text{ID}} = \{(x_i, y_i) | y_i \in \mathcal{Y}_{\text{ID}}\}_{i=1}^{N_{\text{ID}}}$ the marginal distribution over $\mathcal{X} \times \mathcal{Y}$, which represents the distribution of ID data. Suppose a pre-trained model $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}_{\text{ID}}|}$ (with logit outputs) is trained on the training dataset. The task of detecting an OOD instance $x \in \mathcal{X}_{\text{OOD}}$ is to design a decision function $\Psi : \mathcal{X} \to \{0, 1\}$ that distinguishes between ID and OOD for a test input:

$$\Psi(x; f) = \begin{cases} 1 & \text{if } x \in \mathcal{X}_{\text{ID}}, \\ 0 & \text{if } x \in \mathcal{X}_{\text{OOD}}, \end{cases} \tag{1}$$

where Ψ can be defined by a scoring function (e.g., MSP, and Energy) or a parametrized anomaly detector Ψ_θ that is trained with a subset of auxiliary data $\mathcal{D}_{\text{OOD}} = \{(x_i, y_i) | y_i \notin \mathcal{Y}_{\text{ID}}\}_{i=1}^{N_{\text{OOD}}}$. Here, we focus on the OOD detection task by leveraging an OOD detector Ψ_θ.

Synthetic Outlier Generation. Recently, Kim et al. [19] propose KIRBY which constructs a synthetic OOD dataset by replacing class-discriminative features of training samples with marginal background features. We briefly review the KIRBY which is the primary motivation of the proposed method.

KIRBY generates a surrogate OOD sample $\tilde{x}$ by masking class-specific key regions: $\tilde{x} = x_{\text{train}} \odot \mathbf{M}$, where x_{train} is a training sample, $\odot$ is element-wise multiplication, and $\mathbf{M} \in \{0, 1\}^{W \times H}$ is a binary mask indicating class-discriminative features. This mask is obtained by thresholding the output of a pixel attribution method [2, 17, 44]:

$$\mathbf{M}_{ij} = \begin{cases} 0 & \text{if } \mathbf{A}_{ij} \geq \lambda, \\ 1 & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathbf{A} \in [0, 1]^{W \times H}$ is a normalized saliency map and λ is the threshold that determines how many key features are erased. Masked regions are then replaced by marginal features using the fast marching (FM)-based inpainting algorithm $\mathcal{F}$ [39].

4 Method

The models such as CLIP and diffusion models become promising due to their impressive performance and generalization capabilities. Leveraging such models, we propose SHIFT which can generate multiple photo-realistic OOD images from a training sample. In the procedure of SHIFT, the pre-trained CLIP$_{\text{seg}}$ [26] detects regions for ID objects based on an ID class-wise text prompt and erases the detected regions. We then inpaint the regions using the pre-trained latent diffusion model (LDM) [32]. The overall process of SHIFT is described in Figure 2.

4.1 Out-of-Distribution Generation

CLIP$_{\text{seg}}$ has a pre-trained CLIP model as its backbone, followed by a transformer-based decoder that enables pixel-wise prediction. After training on a large scale dataset, CLIP$_{\text{seg}}$ can generate a binary segmentation mask for a test image based on a free-text prompt. We first extract a binarized segmentation mask $\mathbf{M}_{ij} \in \mathbb{R}^{W \times H}$ indicating class-specific key regions from a training image x and its text prompt $\mathcal{Q}_{\text{CLIP}}$:

$$\mathbf{M}_{ij} = \begin{cases} 0 & \text{if } \mathbf{P}_{ij} \geq \lambda, \\ 1 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{P}_{ij} = \text{CLIP}_{\text{seg}}(x, \mathcal{Q}_{\text{CLIP}})$ is the output of CLIP$_{\text{seg}}$, λ is a threshold that determines the extent of removed ID regions in the training image. The text prompt means the ID class and has the form of "*A photo of* `[ID class]`". For example, the prompt for a "car" image is "*A photo of a **car***".

Masked regions are then inpainted using LDM$_{\text{inpaint}}$ to generate perceptually realistic OOD sample $\tilde{x}$:

$$\tilde{x} = \text{LDM}_{\text{inpaint}}(x, \mathbf{M}, \mathcal{Q}_{\text{LDM}}), \tag{4}$$

where $\tilde{x}$ is the generated OOD sample, and $\mathcal{Q}_{\text{LDM}} = \{\mathcal{Q}^P_{\text{LDM}}, \mathcal{Q}^N_{\text{LDM}}\}$ is the static text prompts. $\mathcal{Q}^P_{\text{LDM}}$ is the prompt to generate a background image based on x, and $\mathcal{Q}^N_{\text{LDM}}$ is the negative prompt to prevent ID classes from being created. In this work, we set $\mathcal{Q}^P_{\text{LDM}}$ to be "*A photo of background*" and $\mathcal{Q}^N_{\text{LDM}}$ to be a list of ID classes. In particular, LDM$_{\text{inpaint}}$ can generate multiple synthetic samples from an input image using different seeds. We generate K synthetic OOD samples per training sample. Therefore, for N training samples, we collect the surrogate OOD samples generated by the above process, i.e., $\tilde{\mathcal{D}}_{\text{OOD}} = \{\tilde{x}^1_i, \ldots, \tilde{x}^K_i\}^N_{i=1}$.

Although the high-level concept for generating an OOD sample ($x_{\text{ID}} \rightarrow x_{\text{OOD}}$) is similar to KIRBY, our components are significantly different to the previous work, including additional advantages with respect to OOD detection tasks. The quality of an segmentation mask introduced by KIRBY is depending on a pre-trained classifier because gradient-based saliency method (e.g., Grad-CAM [2] and Layer-CAM [17]) are not model-agnostic and is directly affected by the accuracy of the pre-trained classifier. Compared to KIRBY, our OOD construction is independent to the pre-trained classifier by exploiting the foundation models. In addition, the KIRBY's final OOD data inpainted by $\mathcal{F}$ may include artifacts (the qualitative result for KIRBY is presented in the supplementary material), thereby resulting in a sub-optimal performance.

4.2 Rejection Network

The rejection network Ψ_θ is trained to classify between ID and generated OOD samples. This network can be a binary classifier (ID vs. OOD), and the training objective for Ψ_θ

(with sigmoid function) is as follows:

$$\mathcal{L}(x_i) = \begin{cases} -\log(\Psi_\theta(g(x_i))) & \text{if } x_i \sim \mathcal{D}_{\text{ID}}, \\ -\log(1-\Psi_\theta(g(x_i))) & \text{if } x_i \sim \tilde{\mathcal{D}}_{\text{OOD}}, \end{cases} \tag{5}$$

where $g(x_i)$ is the output feature of the pre-trained classifier's penultimate layer.

When training the rejection network, the weights of the pre-trained network are not updated. Therefore, our training method has the advantage of preserving the ID classification performance of the pre-trained network.

5 Experiments

5.1 Training Details

Pre-trained classifier. We adopt three modern neural networks, including ResNet-34 [8], DenseNet-BC (depth $L = 100$, growth rate $k = 12$) [16], and WideResNet-40-2 [43].
Outlier construction. We adopt the CLIP$_{\text{seg}}$ ("CIDAS/clipseg-rd64-refined") and LDM ("stable-diffusion-2-inpainting") from HuggingFace spaces [40]. Since the foundation models are trained on high-resolution images, an output of the models on an extreme low-resolution image (e.g., 32×32 pixels) may be sub-optimal. We also observe that the OOD samples generated from the original CIFAR dataset are of low quality, and qualitative results are reported in the supplementary material. To handle the low-resolution image, we first refine training images into 512×512 pixel images using the super-resolution diffusion model [32] (the detailed implementation is described in the supplementary material). The converted images are fed into CLIP$_{\text{seg}}$ and the segmentation mask is extracted with λ of 0.2 (Eq. 3). When inpainting the masked image, we set the guidance scale of LDM as 7.5 and we use the default inference steps of 50, alongside the noise schedulers [24]. Lastly, we resize 512×512 OOD images to match the resolution of the ID data.
Rejection network. Ψ_θ is implemented as two fully-connected layers with its hidden layer's width being 2048. Training converges in 10 epochs using SGD-momentum with the initial learning rate of 0.01 and weight decay of 5×10^{-4}.

5.2 Datasets and Metrics

Datasets. Following Kim et al. [19], we compare SHIFT with state-of-the-art methods using CIFAR-10, CIFAR-100 [20], and STL-10 [5] as ID sets and the following six OOD datasets: SVHN [29], DTD [4], Place365 [45], LSUN-crop [23], LSUN-resize [23], iSUN [41].
Metrics. The OOD detction performance is measured with following criteria: (1) **AUROC** is the area under the receiver operating characteristic curve obtained by varying values of the threshold. (2) **FPR@95TPR (FPR)** is the probability that an OOD example is classified as a positive when the true positive rate (TPR) is as high as 95%.

5.3 Baselines

We compare the proposed method with various post-hoc methods: MSP [11], ODIN [23], Mahalanobis [22], Energy [25], ReAct [36], and DICE [35]. Their hyper-parameters are found using grid search based on the respective references and detailed implementations

		Out-of-Distribution Datasets											
	Method	**SVHN**		**DTD**		**LSUN -crop**		**LSUN -resize**		**Place365**		**iSUN**	
		AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓
CIFAR-10 (ID)	MSP	91.91	48.43	88.51	59.11	96.48	25.52	91.07	53.39	89.52	57.04	91.18	50.11
	ODIN	94.70	20.10	88.51	59.11	99.04	4.37	95.13	22.50	91.78	36.63	93.97	28.29
	Mahalanobis	98.58	6.71	96.53	17.76	96.47	22.06	94.98	31.05	82.11	74.05	94.67	30.68
	Energy	91.07	35.35	85.34	52.51	99.05	4.41	93.82	28.91	91.85	34.63	92.24	31.74
	ReAct	90.83	36.81	87.44	51.43	98.91	5.24	93.54	31.39	90.77	35.93	92.19	37.34
	DICE	89.55	36.09	83.35	52.35	**99.59**	1.81	93.87	27.74	90.59	36.65	92.43	33.22
	CSI	95.40	31.92	93.64	64.43	98.68	8.18	98.28	10.23	94.89	31.68	98.30	10.12
	VOS	92.66	32.51	86.58	47.95	99.29	2.94	96.13	20.06	90.70	37.46	95.64	22.62
	GAN	77.15	86.63	72.95	84.87	71.56	88.44	80.92	76.77	80.16	75.68	78.43	80.35
	ACET	91.20	56.21	89.41	51.98	91.55	50.64	91.22	49.54	88.07	55.80	91.30	48.85
	KIRBY	98.99	4.66	95.86	15.84	99.53	2.05	98.66	5.69	95.06	23.05	98.85	4.96
	SHIFT (K=1)	99.31	3.58	95.75	15.84	99.39	2.29	99.26	2.51	96.45	17.61	99.38	2.44
	SHIFT (K=8)	**99.52**	**2.45**	**97.40**	**11.50**	99.57	**1.43**	**99.60**	**1.13**	**96.86**	**15.81**	**99.68**	**0.90**
CIFAR-100 (ID)	MSP	71.37	84.35	73.54	82.65	85.58	60.33	74.11	83.27	70.46	85.17	74.95	83.24
	ODIN	81.34	68.12	76.68	79.53	96.95	16.98	84.96	60.59	72.57	81.74	85.56	59.47
	Mahalanobis	94.66	28.42	90.28	40.05	73.97	76.49	96.08	20.61	66.91	85.83	94.69	25.09
	Energy	73.87	85.61	76.29	79.85	95.88	23.07	77.67	80.94	72.32	82.33	77.93	72.43
	ReAct	87.45	78.05	83.46	68.36	95.06	24.41	72.75	80.76	75.12	78.78	73.41	81.59
	DICE	74.14	86.68	76.65	76.64	**97.84**	11.70	78.84	77.78	73.34	80.60	78.89	79.11
	CSI	85.44	59.49	74.02	73.90	90.19	41.19	92.13	36.10	76.04	75.11	90.59	42.01
	VOS	84.54	75.34	76.56	81.67	97.14	16.59	74.23	80.02	73.19	82.15	73.20	82.68
	GAN	76.07	85.75	66.23	91.50	6 3.99	92.25	64.62	90.08	62.84	91.16	62.06	91.09
	ACET	81.40	73.93	76.19	80.39	77.28	79.68	72.49	78.98	72.53	82.33	73.79	78.45
	KIRBY	96.26	14.96	91.30	**32.43**	97.46	12.50	97.24	14.58	78.94	72.72	97.48	13.03
	SHIFT (K=1)	98.91	4.56	88.42	38.33	97.59	**11.65**	99.82	0.48	84.41	62.64	99.72	1.12
	SHIFT (K=8)	**99.19**	**3.85**	**91.63**	35.80	97.43	13.13	**99.85**	**0.33**	**85.83**	**62.14**	**99.75**	**0.89**
STL-10 (ID)	MSP	57.56	95.98	64.53	89.23	81.64	75.59	80.91	77.07	80.15	77.84	81.60	76.57
	ODIN	98.78	5.20	65.41	83.00	89.17	54.79	87.52	60.27	86.38	72.79	85.70	64.89
	Mahalanobis	98.52	5.13	90.38	**32.25**	70.08	88.90	70.36	87.56	67.88	89.03	84.05	70.19
	Energy	64.03	89.60	64.44	87.71	89.05	54.90	87.71	59.47	86.66	60.57	85.78	63.98
	ReAct	66.48	90.06	71.21	86.73	88.14	56.09	86.99	60.24	86.09	61.55	87.56	63.40
	DICE	70.99	83.91	66.32	82.96	91.40	44.20	89.82	50.68	88.53	53.72	86.30	59.94
	CSI	65.86	90.69	77.70	68.06	83.28	78.37	80.92	72.74	73.78	79.59	82.71	67.85
	VOS	81.36	83.10	66.17	85.40	**92.99**	42.70	83.74	68.26	87.69	59.47	83.12	67.77
	GAN	53.46	97.16	56.52	95.08	60.10	91.98	62.64	90.46	63.08	89.37	62.13	86.95
	ACET	54.15	95.50	65.13	89.14	77.25	79.59	79.35	76.57	79.11	76.38	63.23	90.65
	KIRBY	98.90	1.01	69.64	76.29	90.58	47.16	91.30	43.59	90.31	44.31	92.01	36.36
	SHIFT (K=1)	**99.85**	0.28	86.39	41.49	92.92	**39.70**	93.64	**35.58**	92.17	**40.15**	96.10	21.98
	SHIFT (K=8)	99.84	**0.14**	**91.01**	34.38	92.73	45.30	**93.89**	38.18	**92.84**	41.94	**97.99**	**9.84**

Table 1: Comparison of OOD detection results with state-of-the-art methods using WideResNet. The best and second best results are highlighted in bold and underline, respectively.

are reported in the supplementary material. We compare SHIFT with VOS [6], GAN [21], ACET [10], and KIRBY [19] which learn a rejection rule from OOD samples as described earlier. Lastly, as the orthogonal research, contrastive learning method (CSI [37]) that efficiently learn informative feature representations is compared. Excluding post-hoc methods, we report averaged AUROC and FPR over five runs.

5.4 Comparison with Baselines

The overall performances of the OOD detection are presented in Table 1. SHIFT outperforms all considered baselines on most ID and OOD pairs, even though the rejection network is not exposed to a real OOD dataset. KIRBY and SHIFT generally perform much better than the other methods, but SHIFT shows state-of-the-art performance on the harder STL-10 which has the small amount of the training samples (5,000 images) and its higher resolution (96×96 pixels). One of the reasons for the performance improvement is that SHIFT does not depend on the size of the training dataset compared to KIRBY because it can generate multiple outliers from a single image. SHIFT (K=1) also outperforms the baselines, showing that it can effectively detect outliers even with a small amount of OOD data. The superiority of the SHIFT is again confirmed in Table 2, where the algorithms are further compared when

	WideResNet			**ResNet**			**DenseNet**		
ID	CIFAR10	CIFAR100	STL10	CIFAR10	CIFAR100	STL10	CIFAR10	CIFAR100	STL10
MSP	91.18	75.00	74.40	85.62	82.77	74.40	86.48	77.97	50.39
ODIN	93.86	83.01	85.49	86.67	85.14	83.14	87.32	81.83	52.36
Energy	92.24	78.99	79.61	85.63	85.23	79.19	85.75	80.97	50.42
ReAct	92.27	81.19	81.08	85.63	85.42	80.20	86.64	82.39	50.07
Mahalanobis	93.89	86.10	80.21	95.11	82.97	66.98	90.65	84.21	74.14
DICE	91.56	79.95	82.23	90.68	72.91	83.65	93.53	86.09	84.83
CSI	96.53	84.73	77.37	97.30	92.44	89.16	89.02	73.54	62.73
VOS	93.50	79.81	82.51	93.19	79.04	79.02	95.25	78.78	64.88
GAN	76.86	65.97	59.66	78.23	70.11	61.10	75.32	66.81	59.32
ACET	90.46	75.61	69.70	91.05	74.21	69.12	92.16	77.14	69.98
KIRBY	97.82	93.11	88.79	97.51	90.44	82.30	96.54	92.99	88.59
SHIFT (K=8)	**98.76**	**95.61**	**94.72**	**98.68**	**95.24**	**95.26**	**99.02**	**96.27**	**96.69**

Table 2: The detection performance using different classifier architectures. Each value is the averaged AUROC over the six OOD benchmark datasets.

ID	**Method**	**OOD Datasets**						
		SVHN	**Textures**	**LSUN-crop**	**LSUN-resize**	**Place-365**	**iSUN**	**Average**
CIFAR10	OE	98.36	**97.77**	**99.68**	98.88	96.58	98.79	98.34
	SHIFT (K=8)	**99.52**	97.40	99.57	**99.60**	**96.86**	**99.68**	**98.76**
CIFAR100	OE	87.66	84.39	97.38	78.53	81.93	77.74	84.60
	SHIFT (K=8)	**99.19**	**91.63**	**97.43**	**99.85**	**85.83**	**99.75**	**95.61**

Table 3: Comparison between SHIFT and OE when using WideResNet classifier. Each value is the AUROC. Recall that OE's auxiliary dataset is limited to 32×32 images, and its method naturally does not scale to larger images without additional synthetic constructions.

using different pre-trained classifiers.

5.5 Comparison with Outlier Exposure

OE, unlike all the other algorithms of synthetic OOD set construction, relies on a much larger external OOD set for its reject class. To evaluate whether the OOD data we generated are indeed valuable, we compare SHIFT with OE. In Table 3, we observe that SHIFT (K=8) is shown to be better detection performance on both CIFAR-10 and CIFAR-100. In addition, our method preserve the ID classification accuracy of the pre-trained classification network's because we do not fine-tune the network's parameters. However, OE requires fine-tuning, and its ID accuracy decreases from 94.84 → 94.80 and 75.96 → 75.62 for CIFAR-10 and CIFAR-100, respectively [12].

5.6 Ablation Study

In Table 4, we assess how the choice of ID region removal algorithms affects OOD detection performance by experimenting with Layer-CAM [17] and CLIP$_{\text{seg}}$. Following KIRBY, we extract the saliency map of Layer-CAM at the layer preceding global averaged pooling layer and we then erase the ID key features using the saliency mask (Eq. 2). Compared to Layer-CAM, the OOD sample generated by CLIP$_{\text{seg}}$ that is trained with pixel-level supervision for large scale images more contributes to improving the detection performance of the rejection network. Meanwhile, the OOD detection performance is advanced when the inpainting component is applied.

Layer-CAM	**CLIP$_{\text{seg}}$**	**LDM**	**CIFAR-10** AUROC↑ / FPR↓	**CIFAR-100** AUROC↑ / FPR↓	**STL-10** AUROC↑ / FPR↓
✓			95.30 / 19.70	88.81 / 29.04	84.28 / 58.27
	✓		96.40 / 16.27	91.89 / 24.77	86.63 / 53.22
	✓	✓	**98.27 / 7.33**	**94.85 / 19.20**	**93.47 / 30.02**

Table 4: Ablation study assessing each component in SHIFT (K=1) using WideResNet. We report AUROC and FPR averaged over OOD test sets.

FID↓ / Hausdorff↓	**CIFAR-10** (train)	**CIFAR-100** (train)	**STL-10** (train)
KIRBY	230.9/5.4	206.8/17.7	262.3/12.8
SHIFT (K=1)	**66.4/3.7**	**37.4/14.5**	**114.1/8.0**
CIFAR10 (test)	3.2/3.1	35.8/21.8	97.3/12.7
CIFAR100 (test)	35.7/3.9	3.6/21.6	127.7/12.7
STL10 (test)	93.4/6.1	124.4/22.1	6.8/12.6

Table 5: FID and Hausdorff distance between ID and OOD data. We calculate the Hausdorff distance at the penultimate layer of pre-trained WideResNet.

One of the advantages of LDM is the diversity of its generation. To effectively utilize this characteristic, we generate multiple OOD samples from a single ID sample using different seeds of LDM. To investigate its effectiveness, we trained a rejection network with progressively increasing numbers of K (Figure 3). SHIFT shows gradually improved AUROC as K increases, with optimal performance convergence observed within a range of K between 7 and 9. The diversity afforded by the LDM confers a notable advantage in training the rejection network, as it facilitates the provision of various OOD samples.

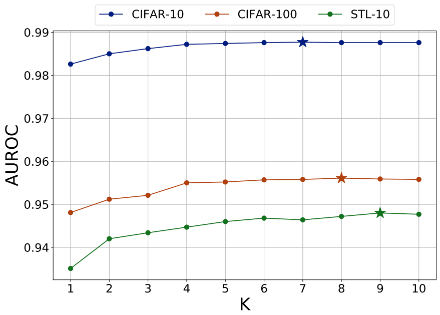

Figure 3: The OOD detection performance (AUROC) with varying K. Each $\star$ marker denotes the point at which the AUROC is highest.

5.7 Analysis

To assess the quality of synthetic images generated by SHIFT, we calculate the Frechet Inception Distance (FID) between ID and OOD samples (Table 5). SHIFT has a higher similarity between the distribution of the generated image and the distribution of the ID image than KIRBY, and this result shows that SHIFT can produce realistic OOD images.

Another valuable characteristic of SHIFT is that the surrogate OOD samples have similar representations to ID samples. For example, in the feature space of WideResNet trained on CIFAR-10, the Hausdorff distance between the SHIFT's OOD and ID samples has the closest distance compared to other OOD datasets. Several studies have been observed that synthetic OOD data is most effective when nearby ID data [19, 21] because excessively distant synthetic examples (easy-to-learn) from ID samples may not help with OOD detection. Furthermore, Ming et al. [28] demonstrate that OOD images that do not have identity objects present, but only have a background similar to the training dataset, lie near the discriminative boundaries of the pre-trained classifier. Relatedly, the OOD samples produced by SHIFT are close to ID data because it generates OOD samples by erasing only ID objects.

6 Conclusion

We proposed the effective OOD generator using vision foundation models where SHIFT removes class-discriminative regions using the CLIP and LDM fills the erased features. Its resultant surrogate OOD dataset is realistic and close to the distribution of ID samples on the feature space of the pre-trained classifier. We demonstrate that these synthetic OOD samples are indeed useful for enhancing the OOD detection performance of the rejection network.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[3] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 430–445. Springer, 2021.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[5] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[6] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022.

[7] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 3266–3273. IEEE, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[10] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[12] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[17] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps. *IEEE Transactions on Image Processing*, 2021.

[18] Jaeyoung Kim, Hong-Seok Lee, In-Seok Song, and Kyu-Hwan Jung. Dentnet: Deep neural transfer network for the detection of periodontal bone loss using panoramic dental radiographs. *Scientific reports*, 9(1):17615, 2019.

[19] Jaeyoung Kim, Seo Taek Kong, Dongbin Na, and Kyu-Hwan Jung. Key feature replacement of in-distribution samples for out-of-distribution detection. *arXiv preprint arXiv:2301.13012*, 2022.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Master's thesis, University of Toronto, Department of Computer Science, 2009.

[21] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations*, 2018.

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[23] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

[24] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PlKWVd2yBkY.

[25] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

[26] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.

[27] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR, 2022.

[28] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10051–10059, 2022.

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *In NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[33] Jaemin Son, Joo Young Shin, Hoon Dong Kim, Kyu-Hwan Jung, Kyu Hyung Park, and Sang Jun Park. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, 127(1):85–94, 2020.

[34] Jaemin Son, Joo Young Shin, Seo Taek Kong, Jeonghyuk Park, Gitaek Kwon, Hoon Dong Kim, Kyu Hyung Park, Kyu-Hwan Jung, and Sang Jun Park. An interpretable and interactive deep learning algorithm for a clinically applicable retinal fundus diagnosis system by modelling finding-disease relationship. *Scientific Reports*, 13 (1):5934, 2023.

[35] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022.

[36] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[38] Keke Tang, Dingruibo Miao, Weilong Peng, Jianpeng Wu, Yawen Shi, Zhaoquan Gu, Zhihong Tian, and Wenping Wang. Codes: Chamfer out-of-distribution examples against overconfidence issue. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1153–1162, 2021.

[39] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[41] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[42] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9518–9526, 2019.

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.