Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need

Jingyao Li[1]   Pengguang Chen[2]   Zexin He[1]   Shaozuo Yu[1]   Shu Liu[2]   Jiaya Jia[1,2]

The Chinese University of Hong Kong[1]   SmartMore[2]
jingyao.li@link.cuhk.edu.hk   leojia@cse.cuhk.edu.hk

Abstract

The core of out-of-distribution (OOD) detection is to learn the in-distribution (ID) representation, which is distinguishable from OOD samples. Previous work applied recognition-based methods to learn the ID features, which tend to learn shortcuts instead of comprehensive representations. In this work, we find surprisingly that simply using reconstruction-based methods could boost the performance of OOD detection significantly. We deeply explore the main contributors of OOD detection and find that reconstruction-based pretext tasks have the potential to provide a generally applicable and efficacious prior, which benefits the model in learning intrinsic data distributions of the ID dataset. Specifically, we take Masked Image Modeling as a pretext task for our OOD detection framework (MOOD). Without bells and whistles, MOOD outperforms previous SOTA of one-class OOD detection by 5.7%, multi-class OOD detection by 3.0%, and near-distribution OOD detection by 2.1%. It even defeats the 10-shot-per-class outlier exposure OOD detection, although we do not include any OOD samples for our detection. Codes are available at https://github.com/lijingyao20010602/MOOD.

1. Introduction

A reliable visual recognition system not only provides correct predictions on known context (also known as in-distribution data) but also detects unknown out-of-distribution (OOD) samples and rejects (or transfers) them to human intervention for safe handling. This motivates applications of outlier detectors before feeding input to the downstream networks, which is the main task of OOD detection, also referred to as novelty or anomaly detection. OOD detection is the task of identifying whether a test sample is drawn far from the in-distribution (ID) data or not. It is at the cornerstone of various safety-critical applications, including medical diagnosis [5], fraud detection [45], autonomous driving [14], etc.
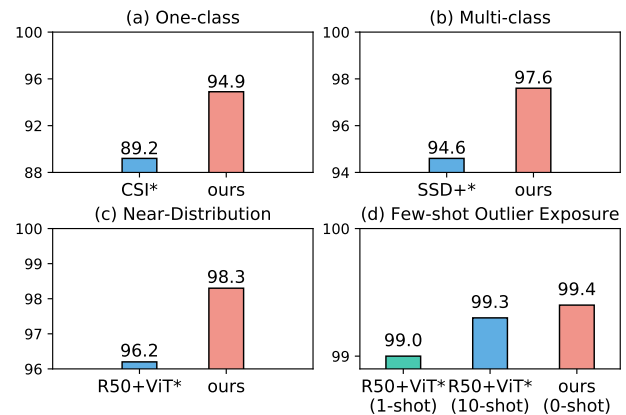

Figure 1. Performance of MOOD compared with current SOTA (indicated by '*') on four OOD detection tasks: (a) one-class OOD detection; (b) multi-class detection; (c) near-distribution detection; and (d) few-shot outlier exposure OOD detection.

Many previous OOD detection approaches depend on outlier exposure [15, 53] to improve the performance of OOD detection, which turns OOD detection into a simple binary classification problem. We claim that the core of OOD detection is, instead, to learn the effective ID representation to discover OOD samples without any known outlier exposure.

In this paper, we first present our surprising finding – that is, *simply using reconstruction-based methods can notably boost the performance on various OOD detection tasks*. Our pioneer work along this line even outperforms previous few-shot outlier exposure OOD detection, albeit we do not include any OOD samples.

Existing methods perform contrastive learning [53,58] or pretrain classification on a large dataset [15] to detect OOD samples. The former methods classify images according to the pseudo labels while the latter classifies images based on ground truth, whose core tasks are both to fulfill the classification target. However, research on backdoor attack [50,51] shows that when learning is represented by classifying data, networks tend to take a shortcut to classify images.

In a typical backdoor attack scene [51], the attacker adds secret triggers on original training images with the visibly correct label. During the course of testing, the victim model classifies images with secret triggers into the wrong category. Research in this area demonstrates that networks only learn specific distinguishable patterns of different categories because it is a shortcut to fulfill the classification requirement.

Nonetheless, learning these patterns is ineffective for OOD detection since the network does not understand the intrinsic data distribution of the ID images. Thus, learning representations by classifying ID data for OOD detection may not be satisfying. For example, when the patterns similar to some ID categories appear in OOD samples, the network could easily interpret these OOD samples as the ID data and classify them into the wrong ID categories.

To remedy this issue, we introduce the reconstruction-based pretext task. Different from contrastive learning in existing OOD detection approaches [53, 58], our method forces the network to achieve the training purpose of reconstructing the image and thus makes it learn pixel-level data distribution.

Specifically, we adopt the masked image modeling (MIM) [2, 11, 20] as our self-supervised pretext task, which has been demonstrated to have great potential in both natural language processing [11] and computer vision [2, 20]. In the MIM task, we split images into patches and randomly mask a proportion of image patches before feeding the corrupted input to the vision transformer. Then we use the tokens from discrete VAE [47] as labels to supervise the network during training. With its procedure, the network learns information from remaining patches to speculate the masked patches and restore tokens of the original image. The reconstruction process enables the model to learn from the prior based on the intrinsic data distribution of images rather than just learning different patterns among categories in the classification process.

In our extensive experiments, it is noteworthy that masked image modeling for OOD detection (MOOD) outperforms the current SOTA on all four tasks of one-class OOD detection, multi-class OOD detection, near-distribution OOD detection, and even few-shot outlier exposure OOD detection, as shown in Fig. 1. A few statistics are the following.

1. For one-class OOD detection (Tab. 6), MOOD boosts the AUROC of current SOTA, i.e., CSI [58], by **5.7%** to **94.9%**.
2. For multi-class OOD detection (Tab. 7), MOOD outperforms current SOTA of SSD+ [53] by **3.0%** and reaches **97.6%**.
3. For near-distribution OOD detection (Tab. 2), AUROC of MOOD achieves **98.3%**, which is **2.1%** higher than the current SOTA of R50+ViT [15].

4. For few-shot outlier exposure OOD detection (Tab. 9), MOOD (**99.41%**) surprisingly defeats current SOTA of R50+ViT [15] (with **99.29%**), which makes use of 10 OOD samples per class. It is notable that we do not even include any OOD samples in MOOD.

2. Related Work

2.1. Out-of-distribution Detection

A straightforward out-of-distribution (OOD) approach is to estimate the in-distribution (ID) density [10, 63, 67, 72] and reject test samples that deviate from the estimated distribution. Alternative methods base on the image reconstruction [1, 17, 33], learn the decision boundary between in- and out-of-distribution data [27,37,68], compute the distance between train and test features [40,53,56,58,59], etc..

In comparison, our work focuses on distance-based methods and yet includes the reconstruction-based methods as a pretext task. The key idea of distance-based approaches is that the OOD samples are supposedly far from the center of the in-distribution (ID) data [65] in the feature space. Representative methods include K-nearest Neighbors [59], prototype-based methods [40, 56], etc.. We will explain the difference between our work and previous OOD detection methods later in this paper.

2.2. Vision Transformer

Transformer has achieved promising performance in computer vision [2, 20] and natural language processing [11]. Existing OOD detection research [15] performs vision transformer (ViT [13]) with classification pre-train on ImageNet-21k [49]. It mainly explores the impact of different structures on OOD detection tasks while we deeply explore the effect from four dimensions for OOD detection, including various pretext tasks, architectures, fine-tune processes, and OOD detection metrics.

It is notable that extra OOD samples are utilized in various previous methods [15, 53] to further improve performance. In contrast, we argue that the exposure of OOD samples violates the original intention of OOD detection. In fact, a sufficient pretext task can achieve comparable or even superior results. Therefore, in our work, we focus on exploring an appropriate pretext task for OOD detection without including any OOD samples.

2.3. Self-Supervised Pretext Task

It has been long in the community to pre-train vision networks in various self-supervised manners, including generative learning [2,11,46,60], contrastive learning [6,7,21,29] and adversarial learning [18, 39, 69]. Among them, representative generative approaches include auto-regressive [46,60], flow-based [12,30], auto-encoding [2,11], and hybrid generative methods [55,66].

| In-Distribution | CIFAR-10 $\longrightarrow$ | | | | CIFAR-100 $\longrightarrow$ | | | |
Out-of-Distribution	SVHN	CIFAR-100	LSUN	Avg	SVHN	CIFAR-10	LSUN	Avg
Classification	98.3	98.6	98.6	98.5	78.0	93.5	88.6	86.7
MoCov3	98.6	92.4	89.8	93.6	78.8	72.8	75.8	75.8
MIM	**99.8**	**99.4**	**99.9**	**99.7**	**96.5**	**98.3**	**96.3**	**97.0**

| In-Distribution | ImageNet-30 $\longrightarrow$ | | | | | | | |
Out-of-Distribution	Dogs	Places365	Flowers102	Pets	Food	Dtd	Caltech256	Avg
Classification	**99.7**	98.4	99.9	**99.6**	**98.3**	98.6	96.8	98.8
MoCov3	88.2	82.0	99.3	81.1	71.4	91.3	88.5	86.0
MIM	99.4	**98.9**	**100.0**	99.1	96.6	**99.5**	**98.9**	**98.9**

Table 1. **Pretext Task**. AUROC (%) of OOD detection on ViT with different pretext tasks on ImageNet22k.

The self-supervised pretext task in our framework is Masked Image Modeling (MIM). It generally belongs to auto-encoding generative approaches. MIM was first proposed in natural language processing [2]. Its language modeling task randomly masks varying percentages of tokens of text and recovers the masked tokens from encoding results of the rest of text. Follow-up research [11, 20] transfers the similar idea from natural language processing to computer vision, masking different proportions of the image patches to recover results.

Multiple existing methods take advantage of self-supervised tasks to guide learning of representation for OOD detection. The latest work [53, 58] presents contrastive learning models as feature extractors. However, existing approaches of classifying transformed images according to contrastive learning possess similar limitations – that is, the model tends to learn the specific patterns of categories, which are beneficial for classification but do not help understand intrinsic data distributions of ID images.

Research of [15] also mentioned this problem. However, the introduced large-scale pre-trained transformers [15] may not jump out of the loop, in our observation, because the pretext task remained to be classification. In our work, we address this issue by performing the masked image modeling task for OOD detection.

3. Method

In this section, we first explain the main factors to help OOD detection and finally propose our framework to achieve this goal.

We first define the notations. For a given dataset X_{ID}, the goal of out-of-distribution (OOD) detection is to model a detector that identifies whether an input image $x \in X_{\text{ID}}$ or $x \notin X_{\text{ID}}$ (that is, $x \in X_{\text{OOD}}$). A majority of existing methods for OOD detection define an OOD score function $s(x)$. Its abnormal high or low value represents that x is from out-of-distribution.

3.1. Choosing the Pretext Task

In this section, we choose the pretext task that can provide the intrinsic prior to suit the OOD detection task. Most previous OOD methods learn the ID representation through classification [15, 23] or contrastive learning [53, 58] on ID samples, which take advantage of either the ground truth or pseudo labels to supervise the classification networks.

On the other hand, work of [50, 51] shows that classification networks only learn different patterns among training categories because it is a shortcut to fulfill classification. It is indicated that the network actually does not understand the intrinsic data distribution of the ID images.

In comparison, the reconstruction-based pretext task forces the network to learn the real data distribution of the ID images during training to reconstruct the image instead of the patterns for classification. Benefiting from these priors, the network can learn a more representative feature of the ID dataset. It enlarges the divergence between the OOD and ID samples.

In our method, we pre-train the model with Masked Image Modeling (MIM) pretext [11] on a large dataset and fine-tune it on the ID dataset. We compare the performance of MIM and contrastive learning pretext task MoCov3 [8] in Tab. 1. It shows that the performance of MIM is much increased by 13.3% to 98.66%.

3.2. Exploring Architecture

To explore an effective architecture [15], we evaluate OOD detection performance on BiT (Big Transfer [31]) and MLP-Mixer, in comparison with ViT. We adopt CIFAR-100 and CIFAR-10 [32] as the ID-OOD pair. They have close distributions because of their similar semantics and construction. Results are in Tab. 2.

R50 + ViT [13, 22] is the current SOTA on near-distribution OOD detection [15], which doubles the model size and testing time but achieves only 96.23% (0.70% higher than ViT). However, MIM on a single ViT significantly improves its AUROC to 98.30% (2.07% higher),

Model	Fine-tuned Test Acc(%)	AUROC(%)
BiT R50 [15]	87.01	81.71
BiT R101×3 [15]	91.55	90.10
ViT [15]	90.95	95.53
MLP-Mixer [15]	90.40	95.31
R50 + ViT (SOTA) [15]	**91.71**	**96.23**

Table 2. **Architecture**. AUROC (%) of OOD detection with various architectures. The last line shows our improvement. The ID and OOD datasets are CIFAR-100 and CIFAR-10, respectively.

| One-Class Dataset | fine-tune | | | AUROC(%) |
	MIM-pt	inter-ft	fine-tune	
CIFAR-10	✓			72.2
	✓	✓		**97.9**
CIFAR-100	✓			66.3
	✓	✓		**96.5**
ImageNet-30	✓			75.2
	✓		✓	**92.0**

Table 3. **Fine-tuning** (One-class). AUROC (%) of OOD detection with different fine-tuning processes on one-class CIFAR-10, CIFAR-100 (super-classes) and ImageNet-30.

without any additional source assumption. It manifests that efficient pretext itself is sufficient for producing distinguishable representation – *there is no need to use a larger model or combination of multiple models* in this regard.

3.3. About Fine-Tuning

One-class Fine-tuning. For one-class OOD detection, we pre-train the MIM model and finely tune it on ImageNet-21k [49], as recommended by BEiT [2]. In particular, when performing one-class OOD detection on ImageNet-30, since we do not include the OOD labels during training, we only pre-train it on ImageNet-21k without intermediate fine-tuning. Therefore, we utilize the label smoothing [57] to help the model learn from the one-class fine-tune task on the ID dataset as

$$y_c^{LS} = y_c(1 - \alpha) + \alpha/N_c, \qquad c = 1, 2, \ldots, N_c \quad (1)$$

where c is the index of category; N_c is the number of classes; and α is the hyperparameter that determines smoothing level. If $\alpha = 0$, we obtain the original one-hot encoded y_c and if $\alpha = 1$, we get the uniform distribution.

Label smoothing was used to address overfitting and overconfidence in normal fine-tuning process. We, instead, find that it can be utilized in one-class fine-tuning. The performance of the model before and after one-class fine-tune is illustrated in Tab. 3. It is clear that the model actually learns information from the one-class fine-tuning operation. This may be counter-intuitive because the labels are equal.

The reason is, due to label smoothing, the loss is larger than 0 and persuades the model to update parameters, although the accuracy reaches 1.

Multi-class Fine-tuning. For multi-class OOD detection, we pre-train the MIM model, intermediately use fine-tuning on ImageNet-21k [49], and apply fine-tuning again on the ID dataset. We perform experiments to validate the effectiveness of each stage in Tab. 5. It proves that all stages contribute well to the performance of OOD detection.

3.4. OOD Detection Metric is Important

Here, we compare the performance of several commonly-used OOD detection metrics, including Softmax [23], Entropy [23], Energy [38], GradNorm [26] and Mahalanobis distance [34]. We perform OOD detection with MIM pretext task with each metric – the results are shown in Tab. 5. They prove that the Mahalanobis distance is a better metric for MOOD.

3.5. Final Algorithm of MOOD

To sum up, in this section, we have explored the effect of contributors to OOD detection, including various pretext tasks, architectures, fine-tuning processes, and OOD detection metrics. In general, we find that the finely tuned MOOD on ViT with Mahalanobis distances achieves the best result. The outstanding performance of MOOD demonstrates that an efficient pretext task itself is sufficient for producing distinguishable representation, and there is no need for a larger model or multi-models.

In Sec. 4, we will show that few-shot outlier exposure utilized in multiple existing OOD detection approaches [15, 53] is also unnecessary. The algorithm of MOOD is shown in the Appendix. It mainly includes the following stages.

1. Pre-train the Masked Image Modeling ViT on ImageNet-21k.
2. Apply intermediate fine-tuning ViT on ImageNet-21k.
3. Apply fine-tuning of pre-trained ViT on the ID dataset.
4. Extract features from the trained ViT and calculate the Mahalanobis distance metric for OOD detection.

4. Experiments

In this section, we compare Masked Image Modeling for OOD detection (MOOD) with current SOTA approaches in one-class OOD detection (Sec. 4.1), multi-class OOD detection (Sec. 4.2), near-distribution OOD detection (Sec. 4.3) and OOD detection with few-shot outlier exposure (Sec. 4.4). Our MOOD outperforms all previous approaches on all four OOD detection tasks significantly.

Experimental Configuration. We report the commonly-used Area Under the Receiver Operating Characteristic Curve (AUROC) as a threshold-free evaluation metric for

finetune			CIFAR-10 $\longrightarrow$				CIFAR-100 $\longrightarrow$			
MIM-pt	inter ft	ft	SVHN	CIFAR-100	LSUN	Avg	SVHN	CIFAR-10	LSUN	Avg
✓			62.2	62.9	98.5	74.5	48.4	42.2	96.0	62.2
✓	✓		89.5	90.0	99.8	93.1	74.3	62.0	**98.3**	68.2
✓		✓	99.1	94.6	97.4	97.0	93.7	83.7	91.4	89.6
✓	✓	✓	**99.8**	**99.4**	**99.9**	**99.7**	**96.5**	**98.3**	96.3	**97.0**

finetune			ImageNet30 $\longrightarrow$							
MIM-pt	inter-ft	ft	Dogs	Places365	Flowers102	Pets	Food	Caltech256	Dtd	Avg
✓			60.2	82.7	28.6	41.9	72.5	42.2	29.4	51.1
✓	✓		**100.0**	97.9	99.9	**99.6**	**97.1**	96.9	98.2	98.2
✓		✓	91.3	97.0	95.1	93.8	99.3	84.0	95.4	92.9
✓	✓	✓	99.4	**98.9**	**100.0**	99.1	96.6	**99.5**	**98.9**	**98.9**

Table 4. **Fine-tuning** (Multi-class). AUROC (%) of OOD detection with different fine-tuning processes on multi-class CIFAR-10, CIFAR-100 and ImageNet-30.

In-Distribution Out-of-Distribution	CIFAR-10 $\longrightarrow$				CIFAR-100 $\longrightarrow$			
	SVHN	CIFAR-100	LSUN	Avg	SVHN	CIFAR-10	LSUN	Avg
Softmax [23]	88.6	85.8	90.7	88.4	81.9	81.1	86.6	83.2
Entropy [23]	**99.9**	97.1	98.1	98.4	93.7	94.1	88.7	92.2
Energy [38]	**99.9**	97.0	97.6	98.2	92.8	93.5	86.1	90.8
GradNorm [26]	99.6	94.3	87.8	93.9	61.6	87.7	38.4	62.6
Distance [34]	99.8	**99.4**	**99.9**	**99.7**	**96.5**	**98.3**	**96.3**	**97.0**

In-Distribution Out-of-Distribution	ImageNet-30 $\longrightarrow$							
	Dogs	Places365	Flowers102	Pets	Food	Dtd	Caltech256	Avg
Softmax [23]	96.7	90.5	89.7	95.0	79.8	90.6	90.1	90.3
Entropy [23]	92.5	87.2	97.5	90.6	69.6	94.9	85.7	88.3
Energy [38]	89.7	82.1	95.8	88.1	67.8	93.1	82.3	85.6
GradNorm [26]	74.8	78.7	92.0	70.6	61.5	90.3	74.3	77.5
Distance [34]	**99.4**	**98.9**	**100.0**	**99.1**	**96.6**	**99.5**	**98.9**	**98.9**

Table 5. **Metric**. AUROC (%) of OOD detection with different metrics on multi-class CIFAR-10, CIFAR-100 and ImageNet-30.

detecting OOD score. We perform experiments on (i) CIFAR-10 [32], which consists of 50,000 training and 10,000 testing images with 10 image classes, (ii) CIFAR-100 [32] and CIFAR-100 (super-classes) [32], which consists of 50,000 training and 10,000 testing images with 100 and 20 (super-classes) image classes. respectively, (iii) ImageNet-30 [49], which contains 39,000 training and 3,000 testing images with 30 image classes, and (iv) ImageNet-1k [49], which contains around 120k and 50k testing images with 1k image classes. More details of training settings are given in the Appendix.

4.1. One-Class OOD Detection

We start with the one-class OOD detection. For a given multi-class dataset of N_c classes, we conduct N_c one-class OOD tasks, where each task regards one of the classes as in-distribution and the remaining classes as out-of-distribution. We run our experiments on three datasets, following prior work [3, 16, 25], of CIFAR-10, CIFAR-100 (super-classes),

and ImageNet-30.

Table 6 summarizes the results, showing that MOOD outperforms current SOTA of CSI [58] on all tested cases significantly. The improvement is of 5.7% to 94.9% on average. The improvement is comparatively smaller on ImageNet-30 Tab. 6c. It is because we do not apply intermediate fine-tuning of the model on ImageNet-30. More details are shown in Sec. 3.3. We provide the class-wise AUROC in the Appendix for detailed exhibition.

4.2. Multi-Class OOD Detection

For multi-class OOD Detection, we assume that ID samples are from a specific multi-class dataset. They are tested on various external datasets as out-of-distribution. We perform MOOD on CIFAR-10, CIFAR-100, ImageNet-30 and ImageNet-1k. For CIFAR-10, We consider CIFAR-100 [32], SVHN [41] and LSUN [36] as OOD datasets. For CIFAR-100, We consider CIFAR-10 [32], SVHN [41] and LSUN [36] as OOD datasets. For ImageNet-30, OOD sam-

Method	Plane	Car	Bird	Cat	Dear	Dog	Frog	Horse	Ship	Truck	Average
OC-SVM [3]	65.6	40.9	65.3	50.1	75.2	51.2	71.8	51.2	67.9	48.5	58.8
DeepSVDD [48]	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
AnoGAN [52]	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
OCGANOCGAN [44]	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
Geom [16]	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0
Rot [25]	71.9	94.5	78.4	70.0	77.2	86.6	81.6	93.7	90.7	88.8	83.3
Rot+Trans [25]	77.5	96.9	87.3	80.9	92.7	90.2	90.9	96.5	95.2	93.3	90.1
GOAD [3]	77.2	96.7	83.3	77.7	87.8	87.8	90.0	96.1	93.8	92.0	88.2
CSI (SOTA) [58]	89.9	99.1	93.1	86.4	93.9	93.2	95.1	98.7	97.9	95.5	94.3
ours	$98.6_{\pm0.4}$	$99.3_{\pm0.5}$	$94.3_{\pm0.6}$	$93.2_{\pm0.5}$	$98.1_{\pm0.6}$	$96.5_{\pm0.4}$	$99.3_{\pm0.2}$	$99.0_{\pm0.1}$	$98.8_{\pm0.1}$	$97.8_{\pm0.4}$	$97.8_{\pm0.4}$
(improve)	+8.7	+0.2	+1.2	+6.8	+4.2	+3.3	+4.2	+0.3	+0.9	+2.3	+3.5

(a) CIFAR-10

Method	AUROC
OC-SVM [3]	63.1
Geom [16]	78.7
Rot [25]	77.7
Rot+Trans [25]	79.8
GOAD [3]	74.5
CSI (SOTA) [58]	89.6
ours	**94.8**
(improve)	+5.2

(b) CIFAR-100

Method	AUROC
Rot [25]	65.3
Rot+Trans [25]	77.9
Rot+Attn [25]	81.6
Rot+Trans+Attn [25]	84.8
Rot+Trans+Attn+Resize [25]	85.7
CSI (SOTA) [58]	91.6
ours	**92.0**
(improve)	+0.4

(c) ImageNet-30

Table 6. **One-class OOD detection.** AUROC (%) of OOD methods on one-class (a) CIFAR-10, (b) CIFAR-100 (super-classes) and (c) ImageNet-30. The reported results on CIFAR-10 are averaged over 3 trials. Subscripts denote standard deviation, and bold ones denote the best results. The last line lists improvement of MOOD over the current SOTA.

ples are from CUB-200 [62], Stanford Dogs [28], Oxford Pets [43], Oxford Flowers [42], Food-101 [4], Places-365 [70], Caltech-256 [19], and Describable Textures Dataset (DTD) [9]. For ImageNet-1k, we utilize non-natural images as OOD datasets, which includes iNatualist [61], SUN [64], places [70], Textures [9].

As shown in Tab. 7, MOOD boosts performance of current SOTA of SSD+ [53] by 3.0% to 97.6% and SOTA of GradNorm [26] by 2.8% to 89.1% on ImageNet-1k. We remark that when detecting hard (i.e., near-distribution) OOD samples on ImageNet30 and Food, MOOD still yields decent performance, while previous methods often fail.

Visualization. In Fig. 2, we illustrate the probability distribution of the test samples according to metrics of three OOD detection approaches: baseline OOD detection [23], SSD+ [53], and MOOD. The baseline OOD detection performs softmax as its OOD detection metric, where ID samples tend to have greater value than OOD samples. MOOD and SSD+ perform the Mahalanobis distance as their metrics.

As shown in the figure, the distance of a majority of testing ID samples to the training data is close to zero, demonstrating a similar representation of training and testing ID samples. In contrast, the distances from most OOD samples

to the training data are much larger, especially on CIFAR-10 and ImageNet-30.

Also, in Fig. 2, we reveal that the difference in the distribution of ID and OOD samples according to MOOD is significantly larger compared with other approaches [23, 53]. It demonstrates that MOOD can separate ID and OOD samples more clearly. In order to illustrate the appearance of images in each ID and OOD dataset, we plot several images as examples with their corresponding distances in the Appendix.

4.3. Near-Distribution OOD Detection

Compared with existing approaches on normal OOD detection tasks, SOTA results of near-distribution OOD detection is much worse – AUROC of some ID-OOD pairs [53, 58] is even lower than 70%. Therefore, improving SOTA for near-OOD detection is essential for the application to work on real-world data.

In Tab. 2, we have compared MOOD with the current SOTA on near-distribution CIFAR10-CIFAR100 (ID-OOD) pair, R50+ViT [15], and MOOD outperforms the latter significantly by 2.07% to 98.30%. In this section, we focus on the hard-detected pairs with similar semantics from Sec. 4.1 and Sec. 4.2.

For one-class OOD detection, we adopt 12 hard-detected

In-Distribution	CIFAR-10 $\longrightarrow$				CIFAR-100 $\longrightarrow$			
Out-of-Distribution	SVHN	CIFAR-100	LSUN	Average	SVHN	CIFAR-10	LSUN	Average
Baseline OOD [23]	88.6	85.8	90.7	88.4	81.9	81.1	86.6	83.2
ODIN [35]	96.4	89.6	-	93.0	60.9	77.9	-	69.4
Mahalanobis [34]	99.4	90.5	-	95.0	94.5	55.3	-	74.9
Residual Flows [71]	99.1	89.4	-	94.3	97.5	77.1	-	87.3
Gram Matrix [54]	99.5	79.0	-	89.3	96.0	67.9	-	82.0
Outlier exposure [24]	98.4	93.3	-	95.9	86.9	75.7	-	81.3
Rotation loss [25]	98.9	90.9	–	94.9	-	-	-	-
Contrastive loss [29]	97.3	88.6	92.8	92.9	95.6	78.3	-	87.0
CSI [58]	97.9	92.2	97.7	95.9	-	-	-	-
SSD+ (SOTA) [53]	**99.9**	93.4	98.4	97.2	**98.2**	78.3	79.8	85.4
ours	99.8$_{\pm0.0}$	**99.4**$_{\pm0.0}$	**99.9**$_{\pm0.0}$	**99.7**	96.5$_{\pm0.6}$	**98.3**$_{\pm0.1}$	**96.3**$_{\pm0.6}$	**97.0**
(improve)	-0.1	+6.0	+1.5	+2.5	-1.7	+20.0	+16.5	+11.6

(a) CIFAR

In-Distribution	ImageNet-30 $\longrightarrow$							
Out-of-Distribution	Dogs	Places365	Flowers102	Pets	Food	Caltech256	DTD	Average
Baseline OOD [23]	96.7	90.5	89.7	95.0	79.8	90.6	90.1	90.3
Contrastive loss [29]	95.6	89.7	92.2	94.2	81.2	90.2	92.1	90.7
CSI (SOTA) [58]	98.3	94.0	96.2	97.4	87.0	93.2	97.4	94.8
ours	**99.4**	**98.9**	**100.0**	**99.1**	**96.6**	**99.5**	**98.9**	**98.9**
(improve)	+0.9	+4.9	+3.8	+1.7	+9.6	+6.3	+1.5	+4.1

(b) ImageNet-30

In-Distribution	ImageNet-1k $\longrightarrow$				
Out-of-Distribution	iNaturalist	SUN	Places	Textures	Average
Baseline OOD [23]	87.6	78.3	76.8	74.5	79.3
ODIN [35]	89.4	83.9	80.7	76.3	82.6
Energy [38]	88.5	85.3	81.4	75.8	82.7
Mahalanobis [34]	46.3	65.2	64.5	72.1	62.0
GradNorm (SOTA) [26]	**90.3**	89.0	84.8	81.1	86.3
ours	86.9	**89.8**	**88.5**	**91.3**	**89.1**
(improve)	-3.4	+0.8	+3.7	+10.2	+2.8

(c) ImageNet-1k

Table 7. **Multi-class OOD detection.** AUROC (%) of OOD detection methods on multi-class CIFAR-10, CIFAR-100, ImageNet-30 and ImageNet-1k. The reported results on CIFAR-10 and CIFAR-100 are averaged over 3 trials. Subscripts denote standard deviation, and bold ones stand for the best results. The last line lists improvement of MOOD over the current SOTA approach.

ID-OOD pairs (AUROC under 90%) from the confusion matrix of current one-class OOD detection SOTA of CSI [58]. The semantics of these ID-OOD pairs are more similar than normal ID-OOD combinations, such as trunk and car, deer and horse, etc., leading to their poor OOD detection performance. As shown in Tab. 8, MOOD significantly boosts the AUROC of current SOTA from 78.7% to 93.9%.

For multi-class OOD detection, we examine the large mistakenly-classified value in the OOD-ID confusion matrix, which represents the number of classifying the OOD image to the category in the ID dataset. For example, when the True-Positive Rate (TPR) is 95%, 48 testing tiger images from CIFAR-100 are classified as cat by the current multi-class OOD detection SOTA method of SSD+ [53],

while only 2 of them are wrongly classified by MOOD. More results are shown in Fig. 3. For the listed 12 ID-OOD pairs, MOOD averagely reduces the number of mistakenly-classified OOD samples notably by 79%.

4.4. OOD Detection with Outlier Exposure

Several representative OOD detection methods [15, 53] utilize OOD samples to improve the performance in extra stages. We note they are not included in our work because we generally believe that exposure of OOD samples violates the original intention of OOD detection.

In Tab. 9, we compare MOOD with current SOTA [15] for near-distribution OOD detection with up to 10 OOD samples per class. We surprisingly find that MOOD works

(a) ID: CIFAR-10

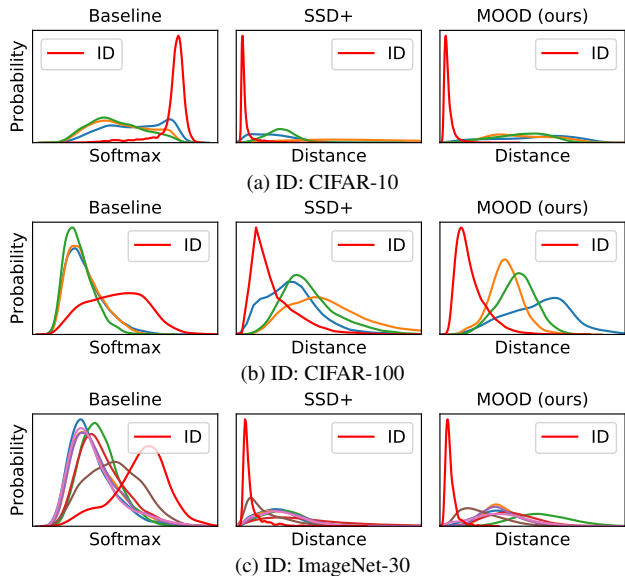(b) ID: CIFAR-100

(c) ID: ImageNet-30

Figure 2. Line chart to illustrate the relation between the probability distribution of test samples and OOD detection metrics on (a) CIFAR-10, (b) CIFAR-100, and (c) ImageNet-30. Each line in the sub-figures represents an OOD or ID dataset. We compare three OOD detection approaches, including baseline OOD detection, SSD+ (current SOTA, [53]), and our proposed MOOD. The baseline OOD detection takes the maximum softmax probabilities as its OOD detection metric, while SSD+ and MOOD both use the Mahalanobis distance as their metrics.

ID class	OOD class	AUROC (%)		
		CSI [58]	ours	(improve)
Plane	Automobile	74.1	99.0	+24.9
Plane	Ship	79.6	99.4	+19.8
Plane	Truck	82.8	98.5	+15.7
Bird	Horse	83.2	94.3	+11.1
Cat	Deer	83.3	92.6	+9.3
Cat	Dog	67.0	75.5	+8.5
Cat	Frog	89.6	92.5	+2.9
Cat	Horse	79.0	95.5	+16.5
Deer	Horse	69.0	100.0	+31.0
Dog	Deer	88.1	96.4	+8.3
Dog	Horse	76.6	95.5	+18.9
Trunk	Automobile	72.3	87.8	+15.5
Average		78.7	93.9	+15.2

Table 8. **Near-distribution OOD detection** (one-class). AUROC (%) of near-distribution pairs in one-class detection on CIFAR-10, compared with current SOTA (CSI [58]).

better in terms of AUROC than current SOTA [15], even though we do not include any OOD samples for detection. The outstanding performance of MOOD demonstrates that an effective pretext task is already sufficient for producing a distinguishable representation that OOD detection requires. Thus, there is no need to include extra OOD samples.
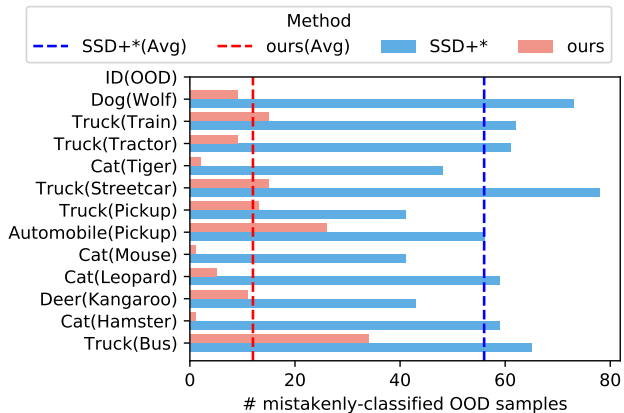

Figure 3. **Near-distribution OOD detection** (multi-class). Number of some mistakenly-classified OOD samples (when TPR = 95%). These samples are wrongly taken as ID samples by the current SOTA of SSD+ [53] in multi-class detection on CIFAR-10. '*' indicates SOTA.

Method	# OOD samples per class	AUROC(%)
R50+ViT (SOTA) [15]	0	98.52
	1	98.96
	2	99.11
	3	99.17
	10	99.29
ours (improve)	0	**99.41**
	-	+0.12

Table 9. **Outlier Exposure OOD detection.** AUROC (%) of current SOTA of R50+ViT [15] for near-distribution OOD detection and MOOD. SOTA utilizes up to 10 known OOD samples per class for detection, while ours do not include any OOD samples.

5. Conclusion

In this paper, we have extensively explored the effect of multiple contributors for OOD detection and observed that reconstruction-based pretext tasks have the potential to provide effective priors for OOD detection to learn the real data distribution of the ID dataset. Specifically, we take the Masked Image Modeling pretext task for our OOD detection framework (MOOD). We perform MOOD on one-class OOD detection, multi-class OOD detection, near-distribution OOD detection, and few-shot outlier exposure OOD detection – MOOD all achieve new SOTA results, although we do not include any OOD samples for detection.

6. Acknowledgement

References

[1] Amir Adler, Michael Elad, Yacov Hel-Or, and Ehud Rivlin. Sparse coding with anomaly detection. *Journal of Signal Processing Systems*, 79(2):179–188, 2015. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3, 4

[3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020. 5, 6

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 6

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021. 3

[9] 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. 6

[10] Gaudenz Danuser and Markus Stricker. Parametric model fitting: From inlier characterization to outlier detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):263–280, 1998. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[12] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 1

[15] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 1, 2, 3, 4, 6, 7, 8

[16] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018. 5, 6

[17] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[22] 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. 3

[23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 3, 4, 5, 6, 7

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 7

[25] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 5, 6, 7

[26] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 4, 5, 6, 7

[27] Nathalie Japkowicz. *Concept learning in the absence of counterexamples: An autoassociation-based approach to classification*. Rutgers The State University of New Jersey-New Brunswick, 1999. 2

[28] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011. 6

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 2, 7

[30] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2

[31] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020. 3

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3, 5

[33] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision*, pages 206–226. Springer, 2020. 2

[34] 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. 4, 5, 7

[35] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 7

[36] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 5

[37] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. 2

[38] 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. 4, 5, 7

[39] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 2

[40] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet*, volume 7, page 9, 2007. 2

[41] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[42] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 6

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[44] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019. 6

[45] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010. 1

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 2

[48] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 6

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 2, 4, 5

[50] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11957–11965, Apr. 2020. 1, 3

[51] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020. 1, 2, 3

[52] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 6

[53] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 1, 2, 3, 4, 6, 7, 8

[54] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv e-prints*, pages arXiv–1912, 2019. 7

[55] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382*, 2020. 2

[56] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Unsupervised clustering approach for network anomaly detection. In *International conference on networked digital technologies*, pages 135–145. Springer, 2012. 2

[57] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 4

[58] 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. 1, 2, 3, 5, 6, 7, 8

[59] Jing Tian, Michael H Azarian, and Michael Pecht. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *PHM Society European Conference*, volume 2, 2014. 2

[60] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[61] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 6

[62] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[63] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 2

[64] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[65] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[66] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 2

[67] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pages 1100–1109. PMLR, 2016. 2

[68] Bangzuo Zhang and Wanli Zuo. Learning from positive and unlabeled examples: A survey. In *2008 International Symposiums on Information Processing*, pages 650–654. IEEE, 2008. 2

[69] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[70] 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. 6

[71] Ev Zisselman and Aviv Tamar. Deep residual flow for novelty detection. 2020. 7

[72] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2