A. Appendix

This supplementary material includes experimental configurations, tables of the figures, visualizations, etc., which are not included in the main paper due to page limitations.

A.1. Algorithm

The algorithm of MOOD is shown Algorithm 1.

Algorithm 1 Masked Image Modeling Out-of-distribution Detection Algorithm

---

Require: Pre-train set X_P, in-distribution set X_{ID}, test set X_{test}, required True Positive Rate $\eta\%$.
Ensure: Is x_{test} outlier or not? $\forall x_{\text{test}} \in X_{\text{test}}$.

1: Partition X_{ID} into the training set X_{train} and calibration set X_{cal}.

2: Pre-train f_{ViT} on X_P by maximizing

$$\sum_{x \in X_P} \mathbb{E}_M \left[\sum_{i \in M} \log p_{\text{MIM}}(z|x^M) \right]$$

.

3: Intermediately Fine-tune f_{ViT} on X_P by minimizing ▷ Not for one-class OOD detection except on ImageNet-30.

$$L_{\text{interft}} = \sum_{x_p \in X_P} \text{CrossEntropy}(f_{\text{ViT}}(x_p), y_P(x_p))$$

4: Fine-tune f_{ViT} on X_{train} by minimizing ▷ Not for one-class OOD detection on ImageNet-30.

$$L_{\text{train}} = \sum_{x \in X_{\text{train}}} \text{CrossEntropy}(f_{\text{ViT}}(x), y^{LS}(x))$$

where y^{LS} is defined by

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

where c is the index of category; N_c is the number of classes; and α is the hyperparameter that determines smoothing level.

5: $h(x) = f_{\text{ViT}}(x)$ for $x \in X_{\text{train}} \cup X_{\text{test}} \cup X_{\text{cal}}$.

6: Use $h(x)$ to calculate $d(x_{\text{test}})$ for $x_{\text{test}} \in X_{\text{test}}$ and $d(x_{\text{cal}})$ for $x_{\text{cal}} \in X_{\text{cal}}$, where $d(\cdot)$ is defined by

$$d_2(x) = \left[(h(x) - \mu)^T {\sum}^{-1} (h(x) - \mu) \right]$$

where μ and $\sum$ are the mean and covariance of the encoding vectors $h(x)$ of the ID training set X_{train}.

7: Compute threshold T as the η percentile of $d(x_{\text{cal}})$.

8: **if** $d(x_{\text{test}}) > T$ **then**

9: $\quad x_{\text{test}}$ is an outlier.

10: **end if**

A.2. Experimental Configuration

We directly utilize the pre-training model released by BEiT [1], which borrows the tokenizer from OpenAI's DALL-E [3] and learns the image tokenizer via a discrete variational autoencoder. During fine-tuning, we follow BEiT and represent the image as a sequence of discrete tokens obtained by an image tokenizer. we randomly crop and resize images in CIFAR to 224×224. Then we split each 224×224 image into a 14×14 grid of image patches, where each patch is 16×16. The patches are linearly-connected and input to the ViT. Our augmentation policy includes random resized cropping, horizontal flipping, and color jittering. More configuration details in the experiments are shown in Tab. A1.

A.3. Detailed Results of One-class OOD Detection

In this section, we exhibit detailed results of one-class OOD detection. Tab. A2 presents the confusion matrix of AUROC values of our method on one-class CIFAR-10. The results align with the human intuition that 'car' is confused for 'truck' and 'cat' is confused for 'dog.' Tab. A3 shows the AUROC of each ID class on ImageNet-30. Tab. A4 presents the OOD detection results of various methods on one-class CIFAR-100 (super-classes).

Baseline	Patch Size	Embed Dimension	Depth	Number of Heads	MLP Ratio	Input Resolution
ViT-Large	16	1024	24	16	4	224

(a) Configuration of the ViT.

Type	Dataset	Intermediate Fine-Tuned	Learning Rate	Warmup Epochs	Epochs	Update Frequency	Layer Decay	Drop Path	Weight Decay	Batch Size
One-Class	CIFAR	✓	2×10^{-3}	5	90	2	0.85	0.1	0.05	64
	ImageNet	×	2×10^{-3}	5	90	2	0.85	0.1	0.05	64
Multi-Class	CIFAR	✓	2×10^{-5}	5	30	2	0.9	0.4	1×10^{-8}	32
	ImageNet	✓	2×10^{-5}	5	50	2	0.9	0.4	1×10^{-8}	32

(b) Configuration of training. CIFAR represents CIFAR-10 and CIFAR-100, and ImageNet represents ImageNet-30.

Table A1. Experimental Configuration

	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
Plane	-	99.0	99.3	99.6	99.6	99.8	99.8	99.4	94.5	98.5	98.8
Car	99.6	-	100.0	99.9	100.0	100.0	100.0	99.9	99.2	93.1	99.1
Bird	96.0	99.5	-	94.2	83.8	95.7	95.0	94.3	98.7	99.4	95.2
Cat	97.5	98.4	95.7	-	92.6	75.5	92.5	95.5	98.6	98.5	93.9
Deer	99.6	99.9	96.9	97.9	-	98.3	98.8	100.0	100.0	100.0	99.1
Dog	99.8	99.9	99.2	83.3	96.4	-	99.2	95.5	100.0	99.9	97.0
Frog	99.8	99.9	99.4	98.4	99.0	99.5	-	99.8	99.9	99.9	99.5
Horse	99.7	99.8	99.4	99.4	95.6	99.2	99.9	-	99.9	99.8	99.2
Ship	96.3	97.9	99.9	99.8	99.8	99.9	100.0	99.7	-	97.5	99.0
Truck	98.9	87.8	100.0	99.9	100.0	100.0	100.0	99.9	98.8	-	98.4
Mean	98.6	98.0	98.9	96.9	96.3	96.4	98.4	98.2	98.8	98.5	97.9

Table A2. Confusion matrix of AUROC (%) values of MOOD on one-class CIFAR-10. The rows and columns indicate the in-distribution and out-of-distribution classes, and the final column indicates the mean value.

ID class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
AUROC(%)	95.8	99.0	98.1	96.6	90.6	92.9	96.9	92.3	92.8	72.9	91.0	94.5	93.8	97.2	82.7
ID class	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
AUROC(%)	97.2	94.0	82.8	74.4	94.4	89.6	90.4	96.8	94.5	80.8	96.9	96.3	90.7	94.2	98.7

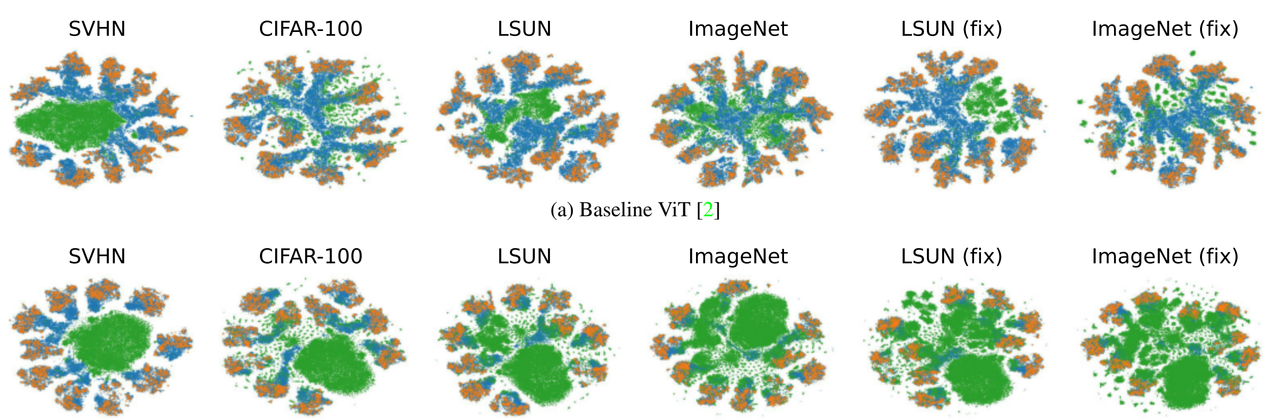
Table A3. AUROC (%) of MOOD on one-class ImageNet-30. The columns indicate in-distribution classes.

A.4. TSNE plot of ViT and MOOD

The t-SNE plot of the features of the baseline ViT [2] and MOOD is shown in Fig. A1. It shows that the OOD samples are classified into ID categories by baseline ViT. In comparison, the OOD samples are gathered tightly and separated from testing samples with MOOD. This visually explains why our framework has a superior capability for OOD detection.

	OC-SVM	DAGMM	DSEBM	ADGAN	Geom	Rot	Rot+Trans	GOAD	CSI	ours
0	68.4	43.4	64.0	63.1	74.7	78.6	79.6	73.9	86.3	**99.5**
1	63.6	49.5	47.9	64.9	68.5	73.4	73.3	69.2	84.8	**94.7**
2	52.0	66.1	53.7	41.3	74.0	70.1	71.3	67.6	88.9	**97.7**
3	64.7	52.6	48.4	50.0	81.0	68.6	73.9	71.8	85.7	**89.5**
4	58.2	56.9	59.7	40.6	78.4	78.7	79.7	72.7	93.7	**96.9**
5	54.9	52.4	46.6	42.8	59.1	69.7	72.6	67.0	81.9	**97.1**
6	57.2	55.0	51.7	51.1	81.8	78.8	85.1	80.0	**91.8**	87.3
7	62.9	52.8	54.8	55.4	65.0	62.5	66.8	59.1	83.9	**97.2**
8	65.6	53.2	66.7	59.2	85.5	84.2	86.0	79.5	91.6	**97.2**
9	74.1	42.5	71.2	62.7	90.6	86.3	87.3	83.7	**95.0**	89.8
10	84.1	52.7	78.3	79.8	87.6	87.1	88.6	84.0	**94.0**	85.1
11	58.0	46.4	62.7	53.7	83.9	76.2	77.1	68.7	90.1	**96.9**
12	68.5	42.7	66.8	58.9	83.2	83.3	84.6	75.1	90.3	**95.4**
13	64.6	45.4	52.6	57.4	58.0	60.7	62.1	56.6	81.5	**97.3**
14	51.2	57.2	44.0	39.4	92.1	87.1	88.0	83.8	**94.4**	93.7
15	62.8	48.8	56.8	55.6	68.3	69.0	71.9	66.9	85.6	**96.7**
16	66.6	54.4	63.1	63.3	73.5	71.7	75.6	67.5	83.0	**93.1**
17	73.7	36.4	73.0	66.7	93.8	92.2	93.5	91.6	**97.5**	95.2
18	52.8	52.4	57.7	44.3	90.7	90.4	91.5	88.0	95.9	**98.7**
19	58.4	50.3	55.5	53.0	85.0	86.5	88.1	82.6	95.2	**97.9**
Mean	63.1	50.6	58.8	55.2	78.7	77.7	79.8	74.5	89.6	**94.8**

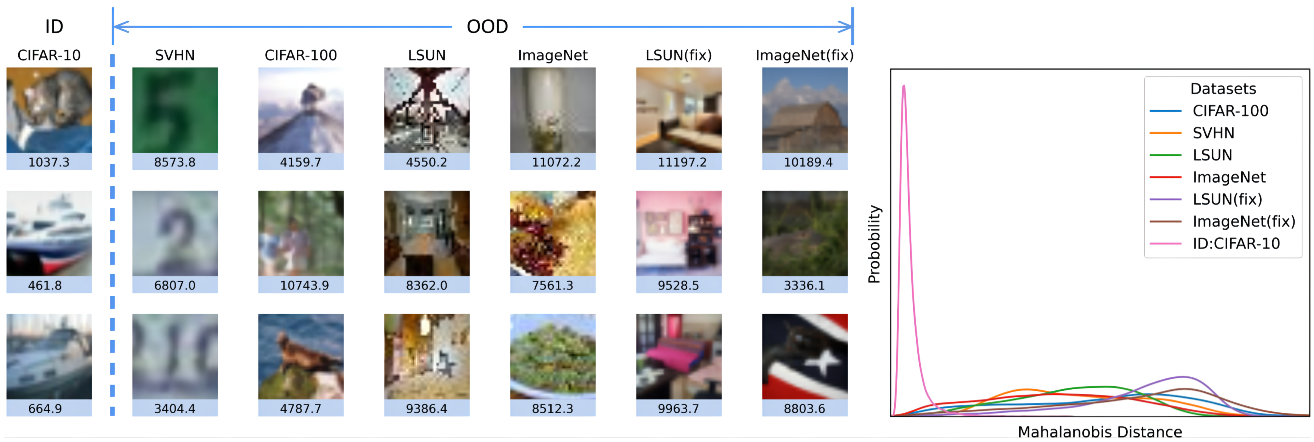
Table A4. AUROC (%) of OOD detection methods on one-class CIFAR-100 (super-classes). The rows and columns indicate the in-distribution classes and OOD detection methods. Bold denotes the best results. The results of previous methods are from the research of [5].



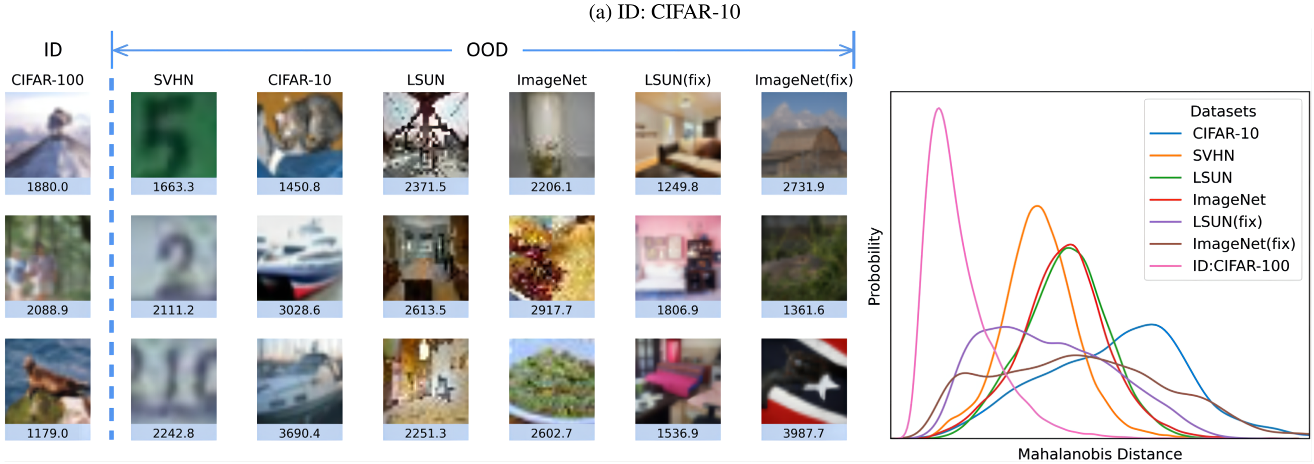
(a) Baseline ViT [2]

(b) MOOD: Masked Image Modeling for OOD

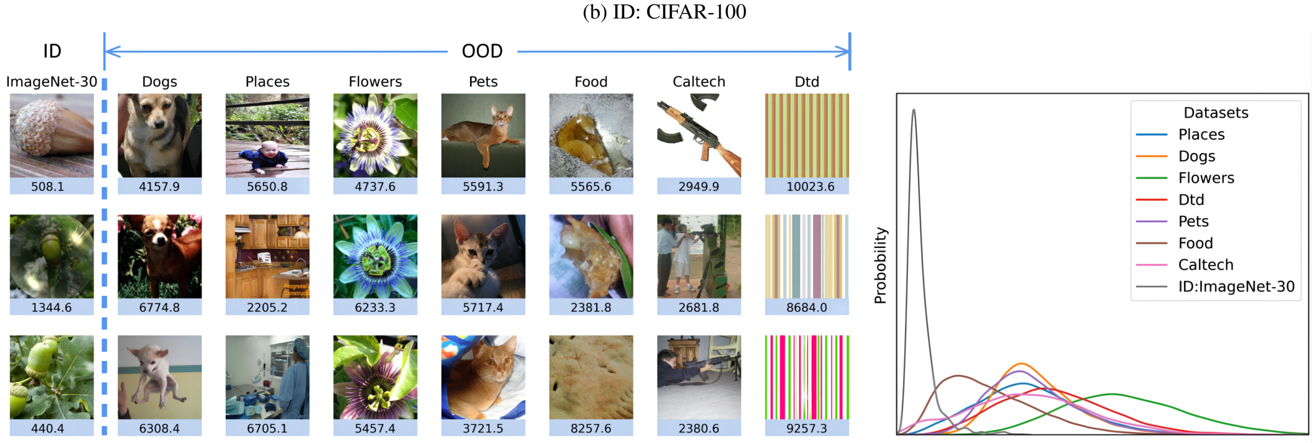
Figure A1. The t-SNE plot of the features on CIFAR-10 of (a) Baseline ViT [2] and (b) MOOD where the subtitles present the out-of-distribution dataset. The three colors represent training, testing and out-of-distribution samples, respectively. It shows that the OOD samples are gathered tightly and separated from testing samples in MOOD, demonstrating its more prominent capability for OOD detection.



(a) ID: CIFAR-10



(b) ID: CIFAR-100



(c) ID: ImageNet-30

Figure A2. We plot the line chart of the distance distribution and some image examples on three ID datasets: (a) CIFAR-10, (b) CIFAR-100, and (c) ImageNet-30. **Line Chart**: The line chart in each sub-figure illustrates the probability distribution of the Mahalanobis distance from the test samples to the mean of training features. Each line represents an OOD or ID dataset. **Images**: We illustrate three images as examples for each ID dataset and its corresponding OOD datasets. The subtitles of the columns of images are the datasets. The first row represents the ID dataset, while the others represent OOD datasets. The corresponding distance of each image is shown below the image in the light blue box.

A.5. Visualization of images

In Fig. A2, we plot the probability distances distribution from the test samples to the mean of training features. The distribution of ID and OOD samples illustrates an obvious gap, which shows that our framework, MOOD, has the potential to distinguish OOD samples from ID data. In order to vividly illustrate the appearance of images in each ID and OOD dataset, we also plot several images as examples with their corresponding distances. For example, in Fig. A2c, the distances of ID images are around 1k, while that of the Describable Textures Dataset (DtD) dataset, which appears to be obviously out-of-distribution, is around 10k.

A.6. Experimental table of mistakenly-classified OOD samples

The mistakenly-classified value in the OOD-ID confusion matrix is shown in Tab. A5, 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 cats by the current multi-class OOD detection SOTA, SSD+ [4], while only 2 of them are wrongly classified by MOOD. For the listed 12 ID-OOD pairs, MOOD averagely reduces the number of mistakenly-classified OOD samples by 79%.

Dataset		# undetected OOD samples		
In-Distribution	Out-Of-Distribution	SSD [4]	MOOD (ours)	(improve)
Truck	Bus	65	34	48%
Cat	Hamster	59	1	98%
Deer	Kangaroo	43	11	74%
Cat	Leopard	59	5	92%
Cat	Mouse	41	1	98%
Automobile	Pickup truck	56	26	54%
Truck	Pickup truck	41	13	68%
Truck	Streetcar	78	15	81%
Cat	Tiger	48	2	96%
Truck	Tractor	61	9	85%
Truck	Train	62	15	76%
Dog	Wolf	73	9	88%
Average		56	12	79%

Table A5. The number of some mistakenly-classified OOD samples (when False-Positive Rate is 95%), that is, classifying to ID category in multi-class detection on CIFAR-10, compared with current SOTA (SSD+ [4]).

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[2] 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

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[4] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 5

[5] 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. 3