Appendix: Out-of-Distribution Detection with Semantic Mismatch under Masking

A Training Process of Generative Model

A.1 Objective Functions

We implement the adversarial loss with a U-net based discriminator [4], denoted as D^{Unet}. D^{Unet} contains two components: D^{Unet}_{enc} and D^{Unet}_{dec}. $D^{Unet}_{enc} \in \mathbb{R}$ provides the real/fake decision as a scalar. While $D^{Unet}_{dec} \in \mathbb{R}^I$ generates a per-pixel real/fake map for the input image, where $I = h \times w$ indicates the scale of input image. Compared with the vanilla discriminator, D^{Unet} not only determines whether the input image is realistic or fake, but also tries to locate the fake parts. Empowered by the per-pixel real/fake map, our generative model can be optimized to focus more on structural semantic features and synthesize coherent image both globally and locally as desired. We formulate the adversarial loss for the discriminator in Eq. (1)-Eq. (3):

$$\mathcal{L}_{D^{Unet}} = \mathcal{L}_{D^{Unet}_{enc}} + \mathcal{L}_{D^{Unet}_{dec}}, \tag{1}$$

$$\mathcal{L}_{D^{Unet}_{enc}} = -\mathbb{E}_x[\log D^{Unet}_{enc}(x, y)] - \mathbb{E}_x[\log(1 - D^{Unet}_{enc}(x', y))], \tag{2}$$

$$\mathcal{L}_{D^{Unet}_{dec}} = -\mathbb{E}_x\Big[\sum_I \log D^{Unet}_{dec}(x, y)\Big] - \mathbb{E}_x\Big[\sum_I \log(1 - D^{Unet}_{dec}(x', y))\Big], \tag{3}$$

where $\mathcal{L}_{D^{Unet}}$ and $\mathcal{L}_{D^{Unet}_{dec}}$ are the loss functions for D^{Unet}_{enc} and D^{Unet}_{dec}, respectively. Correspondingly, the adversarial loss applied on the generator is as follow:

$$\mathcal{L}_{\mathbf{G}} = -\mathbb{E}_x\Big[\log D^{Unet}_{enc}(x', y) + \sum_I \log D^{Unet}_{dec}(x')\Big] + \ell_1(x, x') + \ell_2(x, x') + \mathcal{SSIM}(x, x'). \tag{4}$$

A.2 Training Process

Encoder. We adopt a four-layer convolutional neural network as the feature extractor for Encoder, then two fully-connected layers are employed to output μ and Σ. The dimension of latent variable z is set at 128.

Decoder. We employ the generator architecture proposed in [1] as our Decoder's backbone, then reset the input size to (3, 32, 32), and the channel multiplier to 32, which represents the number of units in each layer [1]. The input latent variable size equals 128.

Discriminator. We build up D^{Unet} based on the implementation of [4], changing the channel multiplier to 32.

All three models mentioned above are trained from scratches in an end-to-end way. We use Adam [3] as the optimizer, with $\beta_1 = 0$, $\beta_2 = 0.999$, learning rate fixed at $5 \cdot 10^{-5}$. The batch size is set at 96. We detail the training process of our generative model in Algorithm 1.

Algorithm 1: Training Framework of **G**

Input : Training data $\mathcal{X} = \{x\}^N$, $\mathcal{Y} = \{y\}^N$, the random mask **M**
Output : The parameters of **E**, **D**

for *some training iterations* **do**
  $x' = \mathbf{G}(\mathbf{M}(x), y) = \mathbf{D}(\mathbf{E}(\mathbf{M}(x), y))$;
  Feed (x, y) and (x', y) into D^{Unet}, respectively;
  Optimize **D** and **E** for $\mathcal{L}_{\mathbf{G}}$(Eq. (4)) and $\mathcal{L}_{KLD}$;
  Optimize D^{Unet} for $\mathcal{L}_{D^{Unet}}$ (Eq. (1));
end
return E, **G**

B Quantitative Results

In this section, we provide more experimental results on CIFAR-10 and CIFAR-100 benchmarks, respectively. In addition, to further validate the effectiveness of the proposed *conditional binary classifier* ($\mathcal{C}_b$) in anomalous scoring model, we detail its performance on each OOD dataset by varying the type of $\mathcal{C}_b$, i.e. trained with/without external OOD data.

B.1 More Results on CIFAR-10 Benchmarks

Table 1 presents the comparison of our MOODCAT trained with external unlabeled data sourcing from TINY-IMAGENET, and baselines implemented with extra data. We conclude that MOODCAT outperforms or at least on par with baselines on CIFAR-10 benchmarks.

Additionally, in Table 1 we observe that OE and UDG achieve a much better performance on SVHN than on other OOD datasets. In fact, most street number images contained in SVHN have relatively flat backgrounds, as shown in Fig. 3 and Fig. 4's SVHN columns. In this case, OE and UDG can achieve excellent performance by overfitting to this specific low-level feature of SVHN instead of considering the semantic level shift caused by SVHN. Thus, when encounter a more challenging case, e.g., CIFAR-100, which has the same data source as CIFAR-10 but different semantic meanings, both OE and UDG suffer a noticeable performance degradation. In contrast, MOODCAT identifies OOD according to their semantic mismatch, thus, remains stable performance on various OODs.

Table 1: OOD Detection Performance on Cifar-10 benchmarks, MoodCat trained with external OOD data. All the values are in percentages. ↑/↓ indicates higher/lower value is better. The best results are in **bold**.

Detection Methods	OOD	FPR@ TPR95% ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑	Classification Accuracy ↑
MCD	Svhn	60.27	89.78	85.33	94.25	90.56
	Cifar-100	74.00	82.78	83.97	79.16	90.56
	Tiny-ImageNet	78.89	80.98	85.63	72.48	87.33
	Texture	83.92	81.59	90.20	63.27	90.56
	Lsun	68.96	84.71	85.74	81.50	90.56
	Places365	72.08	83.51	69.44	92.52	88.51
	Mean	73.02	83.89	83.39	80.53	89.68
OE	Svhn	20.88	96.43	93.62	98.32	91.87
	Cifar-100	58.54	86.22	86.17	84.88	91.87
	Tiny-ImageNet	58.98	87.65	90.09	82.16	89.27
	Texture	51.17	89.56	93.79	81.88	91.87
	Lsun	57.97	86.75	87.69	85.07	91.87
	Places365	55.64	87.00	73.11	94.67	90.99
	Mean	50.53	88.93	87.55	87.83	91.29
UDG	Svhn	**13.26**	**97.49**	**95.66**	**98.69**	92.94
	Cifar-100	47.20	90.98	**91.74**	89.36	92.94
	Tiny-ImageNet	50.18	91.91	**94.43**	86.99	90.22
	Texture	20.43	96.44	98.12	92.91	92.94
	Lsun	42.05	93.21	94.53	91.03	92.94
	Places365	44.22	92.64	87.17	96.66	91.68
	Mean	36.22	93.78	93.61	92.61	92.28
Ours	Svhn	24.27	95.93	92.98	98.05	**95.02**
	Cifar-100	**39.92**	**91.45**	91.54	**91.73**	**95.02**
	Tiny-ImageNet	**32.41**	**93.34**	93.63	**93.41**	**92.54**
	Texture	**6.86**	**98.69**	**99.29**	**97.71**	**95.02**
	Lsun	**33.31**	**93.40**	**93.85**	**93.22**	**95.02**
	Places365	**35.51**	**92.77**	**82.25**	**94.82**	**93.87**
	Mean	**28.71**	**94.27**	**92.26**	**94.82**	**94.42**

B.2 More Results on Cifar-100 benchmarks

Table 2 shows the comparison of our MoodCat trained without external OOD data, and baselines are implemented under the same setting. We conclude that MoodCat achieve the state-of-the-art performance on Cifar-100 benchmarks.

B.3 Ablation Study on Conditional Binary Classifier

To study how much the proposed *Conditional Binary Classifier* ($\mathcal{C}_b$) contributes to MoodCat, we conduct several ablations on $\mathcal{C}_b$. More specific, we consider three configurations: $\mathcal{C}_b$, $\mathcal{C}_b$(Tiny-ImageNet), and $\mathcal{C}_b$+ $\mathcal{C}_b$(Tiny-ImageNet), where $\mathcal{C}_b$ referring to the Conditional Binary Classifier trained only on In-D samples, $\mathcal{C}_b$(Tiny-ImageNet) denoted the Conditional Binary Classifier using Tiny-ImageNet as extra training data and $\mathcal{C}_b$+ $\mathcal{C}_b$(Tiny-ImageNet) indicating that $\mathcal{C}_b$ and $\mathcal{C}_b$ (Tiny-ImageNet) are used in a cascade way.

Table 3 and Table 4 demonstrate $\mathcal{C}_b$'s performance on Cifar-10 and Cifar-100 benchmarks across six OOD datasets, respectively. The main takeaways

Table 2: OOD Detection Performance on CIFAR-100 as In-D, MOODCAT training without external data. All the values are in percentages. ↑/↓ indicates higher/lower value is better. The best results are in **bold**.

Detection Methods	OOD	FPR@ TPR95% ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑	Classification Accuracy ↑
ODIN	SVHN	90.33	75.59	65.25	84.49	76.65
	CIFAR-10	81.28	77.90	79.93	73.39	76.65
	TINY-IMAGENET	82.74	77.58	86.26	61.38	69.56
	TEXTURE	79.47	77.92	86.69	62.97	76.65
	LSUN	80.57	78.22	86.34	63.44	76.10
	PLACES365	76.42	80.66	66.77	89.66	77.56
	Mean	81.89	77.98	78.54	72.56	75.53
EBO	SVHN	78.23	83.57	75.61	90.24	76.65
	CIFAR-10	81.25	78.95	80.01	74.44	76.65
	TINY-IMAGENET	83.32	78.34	87.08	62.13	69.56
	TEXTURE	84.29	76.32	85.87	59.12	76.65
	LSUN	84.51	77.66	86.42	61.40	76.10
	PLACES365	78.37	80.99	68.22	89.60	77.56
	Mean	81.66	79.31	80.54	72.82	75.53
Ours	SVHN	**58.16**	**87.38**	**78.25**	**93.81**	**76.65**
	CIFAR-10	**54.31**	**85.91**	**86.27**	**85.91**	**76.65**
	TINY-IMAGENET	**55.33**	**86.95**	**87.55**	**86.67**	**69.56**
	TEXTURE	**46.70**	**89.20**	**93.48**	**83.28**	**76.65**
	LSUN	**53.43**	**87.98**	**88.82**	**87.32**	**76.10**
	PLACES365	**54.20**	**87.41**	**71.68**	**95.78**	**77.56**
	Mean	**53.69**	**87.47**	**84.34**	**88.80**	**75.53**

are: **(1)** $\mathcal{C}_b$ or $\mathcal{C}_b$(TINY-IMAGENET) alone can achieve acceptable performance; **(2)** $\mathcal{C}_b$(TINY-IMAGENET) outperforms $\mathcal{C}_b$, which means that adding external unlabeled data into the training process can improve the detection ability; **(3)** coupling scorers, here $\mathcal{C}_b$+ $\mathcal{C}_b$(TINY-IMAGENET), usually leads to a better detection capability than that of any single scorer within the coupling. Above findings align with what we have reported in our paper, and further indicate that $\mathcal{C}_b$ plays a key role in the proposed anomalous scoring model.

B.4 Ablation Study on Masking Style

We try several masking forms as exemplified in Fig. 1, and summarize corresponding experimental results in Table 5. Experiments show the randomly masking outperforms other strategies.

From the first three rows in Table 5, we notice that masking can indeed help with performance improvement. However, as we can observed from the second column in Fig. 1, a fixed mask with high ratio (e.g., 0.3) can lead the synthesis to loss fine details. In addition, we implement a patched masking like [2]. However, such masking style may break the continuity within the image, thus lead to low quality on the synthesis for In-D. We also try a non-masking strategy, shuffling, but it further break the continuity of the image. Finally, we identify that the most effective strategy is randomly masking. As can be seen, both the quality for synthesis in Fig. 1 and overall performance in Table 5 outperform other strategies.

Table 3: Conditional Binary Classifier Performance on Cifar-10 benchmarks. All the values are in percentages. ↑/↓ indicates higher/lower value is better. The best results are in **bold**. $\mathcal{C}_b$ and $\mathcal{C}_b$(Tiny-ImageNet) indicates the proposed model trained without/with external unlabeled Tiny-ImageNet data, respectively.

Anomalous Scoring Model	OOD	FPR@ TPR95% ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
$\mathcal{C}_b$	Svhn	48.01	86.85	75.20	94.34
	Cifar-100	42.80	89.13	88.58	89.85
	Tiny-ImageNet	40.54	89.78	89.27	90.42
	Texture	42.54	87.47	91.33	83.85
	Lsun	43.76	90.15	90.18	90.17
	Places365	43.49	89.40	72.82	96.65
	Mean	43.52	88.80	84.56	90.88
$\mathcal{C}_b$ (Tiny-ImageNet)	Svhn	39.47	91.49	83.58	96.23
	Cifar-100	37.43	91.31	91.02	91.73
	Tiny-ImageNet	31.92	93.10	93.01	93.34
	Texture	25.74	94.25	96.17	91.89
	Lsun	32.74	93.55	93.83	93.40
	Places365	34.45	92.78	81.48	97.71
	Mean	33.63	92.75	89.85	94.05
$\mathcal{C}_b$+ $\mathcal{C}_b$ (Tiny-ImageNet)	Svhn	**39.44**	**91.50**	**83.60**	**96.25**
	Cifar-100	**36.64**	**91.40**	**91.15**	**91.85**
	Tiny-ImageNet	**31.86**	**93.12**	**93.04**	**93.38**
	Texture	**25.37**	**94.34**	**96.24**	**92.00**
	Lsun	**32.67**	**93.55**	**93.84**	**93.41**
	Places365	**34.42**	**92.79**	**81.51**	**97.72**
	Mean	**33.40**	**92.78**	**89.90**	**94.10**

B.5 Experiments on Advanced Classifier architectures

We empower UDG with wider (WRN28) and deeper (DenseNet) classifier. Table 6 shows the comparison results with Cifar-100 as In-D samples using WRN28 and DenseNet architecture.

As can be observed from the table, while UDG performs better on these architectures when compared to ResNet18, it still lags far behind our results.

C Qualitative Results

In this section, we demonstrate several batches of visual examples of MoodCat including both In-D and OOD cases.

In-D samples with their syntheses. Fig. 2 visualizes In-D samples and their corresponding syntheses from Cifar-10 and Cifar-100, respectively. Note that we expect the syntheses to resemble the input images for In-D samples with correct labels.

OOD samples with their syntheses. Fig. 3 visualizes OOD samples from six datasets, which are employed in the Cifar-10 benchmarks, along with their corresponding masked images and the syntheses generated by our MoodCat.

Table 4: Conditional Binary Classifier Performance on CIFAR-100 benchmarks. All the values are in percentages. ↑/↓ indicates higher/lower value is better. The best results are in **bold**. $\mathcal{C}_b$ and $\mathcal{C}_b$(TINY-IMAGENET) indicates the proposed model trained without/with external unlabeled TINY-IMAGENET data, respectively.

Anomalous Scoring Model	OOD	FPR@ TPR95% ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
$\mathcal{C}_b$	SVHN	65.18	81.32	65.61	91.35
	CIFAR-10	55.11	85.75	85.78	85.99
	TINY-IMAGENET	54.69	86.27	86.26	86.43
	TEXTURE	56.63	83.30	88.40	77.17
	LSUN	54.77	86.96	87.20	86.83
	PLACES365	54.18	86.36	67.60	95.54
	Mean	56.76	84.99	80.14	87.22
$\mathcal{C}_b$ (TINY-IMAGENET)	SVHN	54.61	**86.30**	**74.30**	93.80
	CIFAR-10	49.82	87.57	87.74	87.69
	TINY-IMAGENET	45.86	89.38	89.48	89.43
	TEXTURE	48.24	87.16	91.55	81.83
	LSUN	44.43	90.07	90.25	90.00
	PLACES365	46.89	88.93	72.99	96.41
	Mean	48.31	88.24	84.39	89.86
$\mathcal{C}_b$+ $\mathcal{C}_b$ (TINY-IMAGENET)	SVHN	**54.31**	**86.30**	**74.30**	**93.81**
	CIFAR-10	**49.62**	**87.60**	**87.77**	**87.77**
	TINY-IMAGENET	**45.46**	**89.39**	**89.48**	**89.48**
	TEXTURE	**47.18**	**87.37**	**91.71**	**82.17**
	LSUN	**44.01**	**90.08**	**90.26**	**90.04**
	PLACES365	**46.73**	**88.95**	**73.02**	**96.43**
	Mean	**47.89**	**88.28**	**84.42**	**89.95**

In Fig. 4, the In-D dataset changes to CIFAR-100. We employed OOD samples sourced from the same six OOD datasets as that of CIFAR-100 benchmarks in Fig. 4. Note that, when OOD is fed to MOODCAT, we prefer to have a clear distinction between the synthesis generated by MOODCAT and the input image.

D Further Discussion

D.1 Computational Cost Analysis

MOODCAT is designed as an auxiliary model, which works in parallel with the classifier. This auxiliary architecture ensures MOODCAT a plug-and-play model without compromising classifier's accuracy. Meanwhile, MOODCAT can satisfy high performance requirements in the context of OOD detection. However, as an auxiliary model, MOODCAT inevitably introduces extra computation and memory costs.

Table 7 summarizes the computational cost of MOODCAT, and that of ODIN, i.e., ResNet18, and that of widely adopted classifier architectures, ResNet18, WResNet28, WResNet101 in terms of number of multiply-add operations (MAC), and number of model's parameters (Params). As can be observed, the cost of basic version of MOODCAT, i.e. $\mathcal{C}_b$, **E**, **D**, is relatively small, Params 4.552M, MACs

Fig. 1: Visualization of different masking styles and their impacts on synthesized images. The semantic label is assigned as "car" for both the In-D image and the OOD image. We set the masking ratio as 0.3 for "Fixed High Ratio" and "Patched", 0.1 for "Fixed Low Ratio", and that of "Randomly" varies from 0.1 to 0.3. MoodCat employs the **Randomly** masking style.

Table 5: Ablation studies on different masking styles. The results are obtained by setting Cifar-10 as In-D, Cifar-100 as OOD, with MoodCat trained on extra Tiny-ImageNet acting as OOD. The **bolded** values are the highest performance. All the values are in percentages. ↑/↓ indicates higher/lower value is better.

Mask Style	FPR@TPR95% ↓	AUROC ↑	AUPR-In ↑	AUPR-Out ↑
Without Masking	40.53	91.26	91.25	91.55
Fixed Low Ratio	40.20	91.33	91.32	91.59
Fixed High Ratio	39.57	91.56	91.48	91.87
Patched	39.81	91.34	91.33	91.64
Shuffling	44.14	88.73	88.15	89.18
Randomly	**39.48**	**91.66**	**91.65**	**91.95**

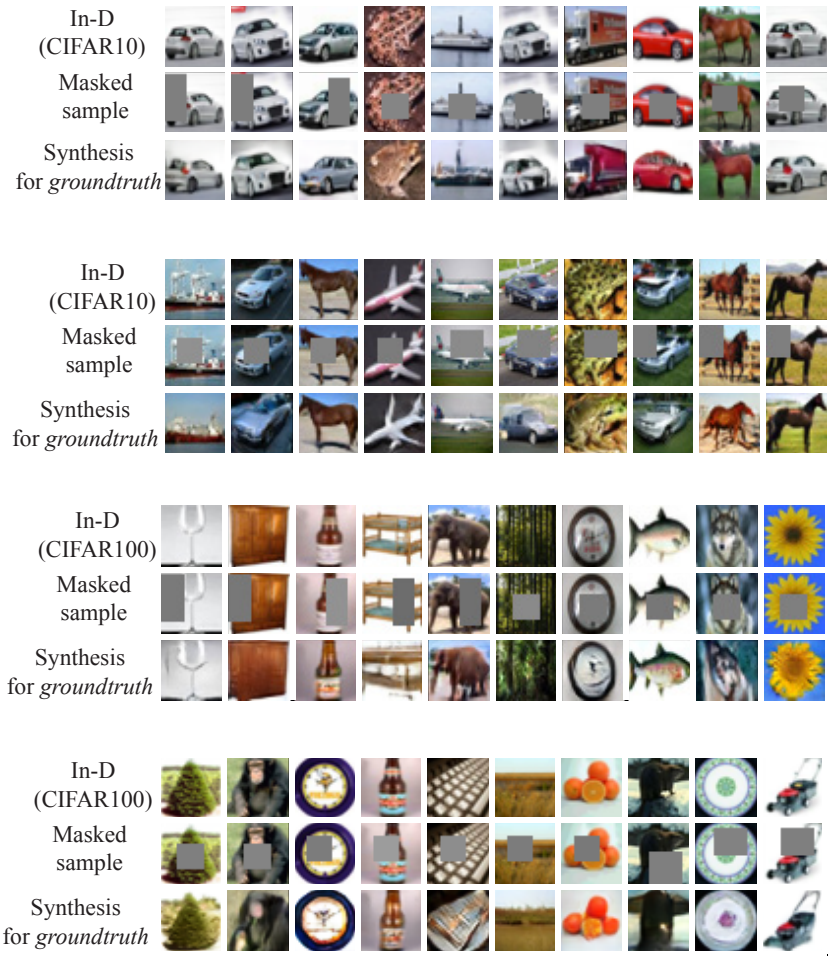

Fig. 2: Visualization results of MoodCat with Cifar-10/ Cifar-100 as In-D. We exemplify several In-D samples in each panel's first row, following the intermediate masked version, and the last row presents their corresponding synthetic version generated by MoodCat with the groundtruth labels.

Fig. 3: OOD visualization results of MoodCat trained on Cifar-10. In each panel, we exemplify OOD samples across six OOD datasets in the first row, following is the intermediate masked version, the last row presents their corresponding synthetic version generated by MoodCat with the given semantic label, the same below.

Table 6: Experiments on advanced model architectures. Performance comparison with UDG on CIFAR-100 benchmarks. For our method, we use the results in the main paper with a ResNet18 classifier. We give advantage to UDG, which is reimplemented with deeper/wider WideResNet-28, DenseNet, while MOODCAT's parameter number is equivalent to ResNet18. **Bold** are the best.

Architecture	OOD dataset	FPR@TPR95 ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
WideResNet28 UDG	SVHN	66.76	85.29	76.14	92.33
	CIFAR-10	82.35	76.67	78.52	72.63
	TINY-IMAGENET	78.91	79.04	87.00	65.06
	TEXTURE	73.62	79.01	85.53	67.08
	LSUN	77.04	79.79	87.49	66.93
	PLACES365	72.25	81.49	66.72	90.65
	Mean±*Std*	75.16±5.49	80.22±2.93	80.23±8.11	75.78±12.44
DenseNet UDG	SVHN	80.67	75.54	75.65	70.99
	CIFAR-10	85.87	74.06	77.16	68.90
	TINY-IMAGENET	82.36	76.81	85.76	61.56
	TEXTURE	76.32	78.93	63.79	89.02
	LSUN	79.12	78.91	66.83	88.23
	PLACES365	73.59	76.27	82.76	65.20
	Mean±*Std*	79.66±4.36	76.75±1.92	75.33±8.64	73.98±11.79
MOODCAT (Ours, Res18)	SVHN	**51.6**	**88.99**	**80.89**	**94.81**
	CIFAR-10	**50.17**	**87.76**	**88.18**	**87.79**
	TINY-IMAGENET	**46.07**	**89.42**	**89.73**	**89.28**
	TEXTURE	**42.22**	**90.56**	**94.43**	**85.13**
	LSUN	**47.85**	**89.96**	**90.33**	**89.23**
	PLACES365	**47.72**	**89.3**	**74.83**	**96.48**
	Mean±*Std*	**47.61**±3.29	**89.33**±9.95	**86.4**±7.19	**90.45**±4.33

0.408G, when compared to that of ResNet18 (Params 11.174M, MAC 0.556G) and other widely adopted classifier architectures, e.g., WResNet28 (Params 36.479M, MACs 5.248G). Note that the performance of basic MOODCAT, whose anomalous scoring model only contains $\mathcal{C}_b$, is still acceptable as shown in Table 3 and Table 4. Thus, if the computational cost is a real concern in the practice, the deployer can adopt MOODCAT with $\mathcal{C}_b$ alone as the anomalous scorer. For the MOODCAT supported by IQA models, e.g., LPIPS, DISTS, the total computational cost is comparable to that of WResNet28 or WResNet101, yet slightly larger than ResNet18. Thus, if the detection ability is put at the first place, one can explore to enhance the anomalous scoring model by employing extra IQA models. Actually, there is a trade-off between the OOD detection performance and computational cost of MOODCAT, and our anomalous scoring model leaves design space for the deployer to explore according to the real-world application.

Table 7: Computational and memory costs of MOODCAT and its components.

Model	**E**	**D**	$\mathcal{C}_b$	**MOODCAT basic**	**LPIP /DISTS**	**MOODCAT**	**ResNet 18**	**WResNet 28**	**WResNet 101**
Params (M)	0.460	3.821	0.271	**4.552**	14.715	33.982	11.174	36.479	126.89
MACs (G)	0.0049	0.297	0.105	**0.408**	0.630	1.718	0.556	5.248	22.84

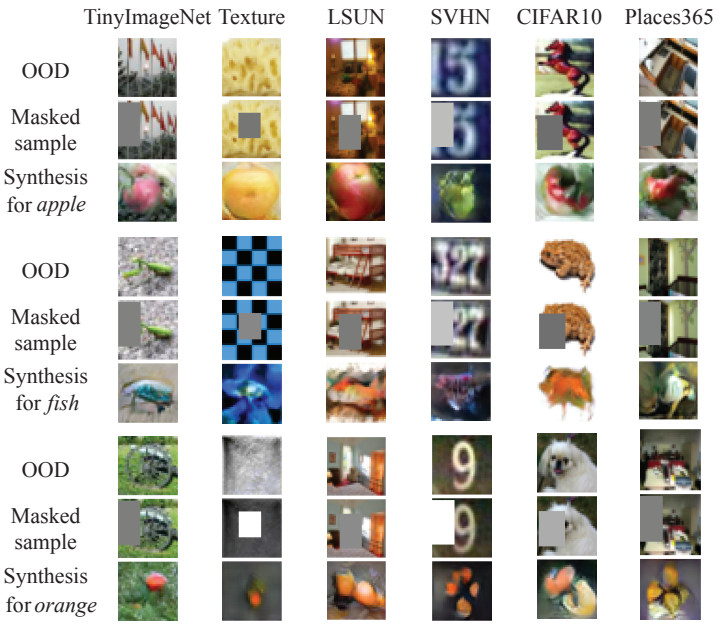

Fig. 4: OOD visualization results of MoodCat trained on Cifar-100.

D.2 Failure Cases

Fig. 5 demonstrates some of MoodCat's failure cases. In Fig. 5 (a), the OOD samples sourcing from Cifar-100, are falsely distinguished as In-D samples (Cifar-10). As can be seen, OODs and their synthetic images resemble to each other for same degree. For example, the first column's "cattle" partly contains some features such as legs and the tail, which match the given semantic label "horse" well, resulting in the synthesis having high image quality while resembling to the input image, therefore leading to the final misjudgement.

Fig. 5 (b) presents several False Negative samples, i.e., samples sourcing from In-D are wrongly predicted as OOD samples. As can be observed, the In-D sample with rare characteristics, e.g. a blue fog, an ostrich with its head down, are more likely to be misclassified as OOD. In addition, if the mask happens to cover the object completely, MoodCat can hardly recover the input image without necessary features, as the cases shown in the third and fourth columns of Fig. 5 (b). Moreover, an poor semantic meaning in the In-D sample itself can lead to the final misclassification. For example, in the last column of Fig. 5 (b), even humans can hardly tell what is depicted in the input image, let alone MoodCat.

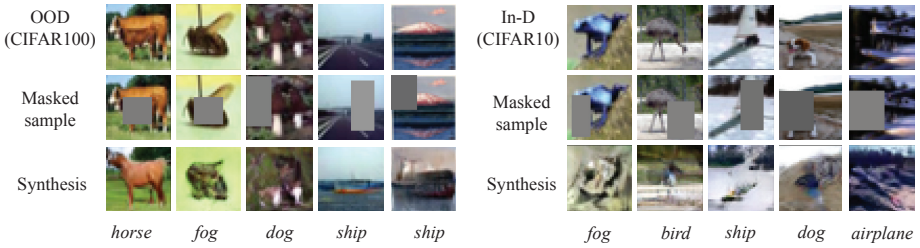

Fig. 5: Failure cases of MoodCat. We exemplify both False Positive and False Negative failure cases in (a) and (b), respectively. (a) False Positive failure cases, where samples come from OOD dataset (Cifar-100) are falsely identified as In-D samples (Cifar-10). (b) False Negative failure cases, where samples belong to In-D are wrongly flagged as OOD samples. The predicted label for each input sample are provided under the corresponding synthetic image.

References

1. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2018)
2. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)
4. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8207–8216 (2020)