A. Additional Training Details

In this section, we provide training details for all models used to report results in section 4 of the main paper.

A.1. Classifier Training

MNIST models: For experiments on MNIST, we use 4 convolutional architectures: LeNet, AlexNet, VGG-11 and ResNet-18. Each model has been trained on a single 12 GB TITAN Xp GPU for 100 epochs using SGD as the optimizer with a momentum of 0.9. The initial learning rate used was 0.1 and there are learning rate drops by a factor of 10 at training epochs 40 and 60. The training batch size used for MNIST is 256.

CIFAR-10/100 models: All convolutional classifiers: DenseNet-121, ResNet-50/110, VGG-16, are trained using the Pytorch framework with a single 12 GB TITAN Xp GPU. To train models on CIFAR-10/100, we use the SGD optimiser with a momentum of 0.9 and a weight decay of $5e^{-4}$. We train each model for 350 epochs using 0.1 as the learning rate and a learning rate drop by a factor of 10 at training epochs 150 and 250. We use a training batch size of 128 and augment the training set using random crops and random horizontal flips.

For Vision Transformer (ViT) models trained on CIFAR-10/100, we use 4 12 GB TITAN Xp GPUs to train a single model. We train 4 different ViT models: ViT-B-16/32 and ViT-L-16/32, using an image size of 224 × 224 and other conventional augmentations including random crop and random horizontal flips. All the ViTs are pretrained on ImageNet-21K. We use a SGD with a momentum of 0.9 and a learning rate of $3e^{-2}$ with a cosine learning rate decay. We use 500 warmup steps for each model and train them for a maximum of 10000 steps. We use a training batch size of 256 for the ViT models.

ImageNet models: For ImageNet, we use pre-trained convolutional models available in the `torchvision.models` library. Furthermore, we use pretrained Vision Transformers for all evaluation purposes.[1]

Classifier Suite for computing $\mathcal{L}_{\text{MI}}$ Note that the in order to compute $\mathcal{L}_{\text{MI}}$, we use a single ensemble containing models, each with a different architecture. For MNIST experiments, we use 4 different models: LeNet, AlexNet, VGG-11 and ResNet-18 (one model of each architecture) as the ensemble to compute mutual information over. Similarly, for CIFAR-10/100, we use 6 models with 6 different architectures: DenseNet-121, ResNet-50/110, VGG-16, Wide-ResNet-28-10 and Inception-v3. Finally, for ImageNet, we use a set of pretrained classifiers from the Pytorch `torchvision.models` library. In particular, we get ResNet-18, MobileNet-v3-Large and EfficientNet-B0. Note that the use of ensembles with different architectures is to encourage higher variability in predictions and representations within the ensemble, thereby encouraging higher mutual information for predictions. Ensembles used for the evaluation of generated samples all have the same architecture. All the classifiers used for computing $\mathcal{L}_{\text{MI}}$ are trained using the same dataset-specific settings as mentioned above.

A.2. Training Pix-2-Pix GAN

In order to train a Pix-2-Pix GAN, we use $\mathcal{L}_{\text{Shift}}$ defined in eq. (3) as the loss function for the generator of the GAN and there is no change to the loss of the discriminator. However, note that the target image for the Pix-2-Pix discriminator is the same as the input. Thus the loss of the discriminator can be given as:

$$\begin{aligned}\mathcal{L}_{\text{D}_{\text{Pix}-2-\text{Pix}}} =& \mathbb{E}_{\mathbf{x}}\left[\log(D(\mathbf{x},\mathbf{x}))\right] - \\ & \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log(1-D(\mathbf{x},G(\mathbf{x},\mathbf{z})))\right]\end{aligned} \tag{5}$$

For $\mathcal{L}_{\text{MI}}$, in eq. (3), we use the method specified above. We use a single 12 GB TITAN Xp GPU to train the Pix-2-Pix model on CIFAR-10 and 8 such GPUs to train on ImageNet. We use a training batch size of 256 and train the model for 100 epochs using Adam as the optimizer, a learning rate of 0.0002 and beta values 0.5 and 0.999. All other training settings are the same as specified in the original Pix-2-Pix paper [30].

A.3. Training GAN

For generating Near OoD samples (i.e., Near OoD), we use a DCGAN for MNIST and a BigGAN for CIFAR-10/100 and ImageNet. We use a single 12 GB TITAN Xp GPU to train DCGAN for MNIST and BigGAN for CIFAR-10/100. However, we use 8 such GPUs to train a single BigGAN on ImageNet. The loss function for the discriminator of the GAN undergoes no change and is shown as follows:

$$\mathcal{L}_{\text{D}_{\text{GAN}}} = \mathbb{E}_{\mathbf{x}}\left[\log D(\mathbf{x})\right] - \mathbb{E}_{\mathbf{z}}\left[\log(1-D(G(z)))\right] \tag{6}$$

The loss function for the generator is $\mathcal{L}_{\text{Near OoD}}$ as shown in eq. (4). The $\mathcal{L}_{\text{MI}}$ in $\mathcal{L}_{\text{Near OoD}}$ is computed as described above. We train all GANs for 100 epochs and all other training details for the GANs are exactly the same as set out in their respective repositories.[2]

B. Outlier Exposure

In [23], exposure to outliers during training was proposed as a way to improve model performance on OoD datasets. In outlier exposure, models are trained on two datasets: i) the training set on which the loss is the usual

[1] See `github.com/rwightman/pytorch-image-models` for details.

[2] See `github.com/ajbrock/BigGAN-PyTorch` for details on training BigGAN.

Model	Outlier Dataset	# Outlier Classes	Test Accuracy	AUROC					
				SVHN	C10	Tiny-ImageNet	N C100	Places365	Texture
ResNet-50	None	0	79.52	80.97	78.98	79.52	56.38	71.17	68.71
	SVHN	10	78.79	–	79.95	80.02	60.45	72.30	70.07
	C10	10	78.98	82.97	–	83.11	62.60	73.23	72.17
	N C100 (Ours)	100	78.82	87.02	81.65	84.40	–	75.80	74.92
	Tiny-ImageNet	200	79.07	86.46	81.57	–	64.11	75.69	74.75
	ImageNet (Subset)	500	78.61	88.17	82.66	85.01	65.32	76.44	74.87
Wide-ResNet-28-10	None	0	80.46	81.46	80.54	81.84	62.69	75.43	73.62
	SVHN	10	80.13	–	82.97	82.77	64.44	76.21	75.13
	C10	10	79.7	84.45	–	82.98	64.52	78.4	77.21
	N C100 (Ours)	100	79.92	87.23	84.31	85.63	–	80.61	80.12
	Tiny-ImageNet	200	79.83	87.09	84.52	–	65.73	80.34	80.06
	ImageNet (Subset)	500	79.74	89.01	85.6	83.57	66.24	81.7	80.77

Table 3: **AUROC% obtained by performing outlier exposure [23] on models trained on CIFAR-100 (C100) with increasingly diverse OoD datasets.** Models tuned using Near OoD CIFAR-100 (N C100) as outliers with 100 outlier classes outperform less diverse outlier sets and perform competitively with more diverse outlier sets including Tiny-ImageNet and our ImageNet subset (500 classes).

cross-entropy and ii) the outlier dataset on which the loss is the cross-entropy between the softmax distribution and a uniform distribution over classes. The assumption is that exposure to good outlier datasets will make the model detect any unseen outlier datasets as well. In this experiment, we want to see how models can improve on OoD detection performance once exposed to increasingly diverse OoD samples as outliers and if our Near OoD samples can be used effectively for outlier exposure.

To do this, we train a ResNet-50 and a Wide-ResNet-28-10 on CIFAR-100 using SVHN, CIFAR-10, Near OoD CIFAR-100, Tiny-ImageNet and a subset of ImageNet as outlier datasets. For the ImageNet outliers, we use a subset of 500 classes from ImageNet which are disjoint from CIFAR-100 and with 100 randomly chosen images from each class. Note that the outlier sets are increasingly diverse with Near OoD CIFAR-100 having the same number of classes as CIFAR-100 and Tiny-ImageNet and the ImageNet subset having 200 and 500 classes respectively. The training procedure is the same as set out in appendix A.1. However, in the loss, following [23], in addition to the cross-entropy term, we also have an additional regulariser which computes the cross-entropy of the output with a uniform distribution for outlier samples. For both architectures, we also compare with a baseline with no exposure to outliers. Finally, we use Places365 [78] Texture [5] as independent OoD datasets which are not used for any outlier exposure. We present the test accuracy and AUROC scores for all models in table 3. *We observe that models trained using Near OoD CIFAR-100 as outliers consistently outperform models trained with relatively less diverse outlier sets like CIFAR-10 and SVHN. Additionally, we find that these models also broadly outperform Tiny-ImageNet and are competitive with models trained using the ImageNet subset which are more diverse outlier sets. All models outperform the ones trained without any outliers.*

The above observation provides additional evidence to support the use of Near OoD samples, not just to benchmark OoD detection baselines, but also to improve them through outlier exposure. It indicates that our generated samples are not obtained from random transformations of in-distribution images but indeed capture desirable properties which make them effective outliers for training more robust models. This corroborates our previous observation that even an image which does not represent any real world object can be very useful if it captures desirable properties in terms of semantic and perceptual similarity.

C. Additional Results

In this section, we present additional results to support the results in the main paper.

MI overlap: In fig. 10, we show the mutual information of the training ensemble on real CIFAR-10 samples along with Near OoD generated samples on CIFAR-10. We choose samples which minimise MI overlap between real and generated samples without having a very high MI as that leads to generated samples losing their perceptual similarity with iD. For CIFAR-10, we choose $[0.2, 0.6]$ as the MI interval for generated samples.

OoD Detection on Near OoD Datasets table 4 presents test set accuracy and AUROC scores of models trained on MNIST on the MNIST vs Fashion-MNIST and MNIST vs Near OoD MNIST. In table 5, we report the CIFAR-10/100 and ImageNet test set accuracy of all the models we use to evaluate our benchmark. In table 11 and table 13, we report the AUROC scores of 6 convolutional models: DenseNet-121, ResNet-50/110, VGG-16, Wide-ResNet-28-10 and Inception-v3 and 4 Vision Transformer models: ViT-B-16/32, ViT-L-16/32 trained on CIFAR-10 and CIFAR-100 respectively. The uncertainty computation method here uses the softmax entropy, softmax confidence and Mahalanobis distance computed from a single deterministic model. We also compute the AUROC scores for a deep ensemble of size 5, using the 6 convolutional architectures and report the correspoding results in table 6 and table 8 for models trained on CIFAR-10 and CIFAR-100 respectively. For CIFAR-10, we use SVHN, CIFAR-100 and Near OoD CIFAR-10 as OoD sets and for CIFAR-100, we use SVHN, CIFAR-10 and Near OoD CIFAR-100 as

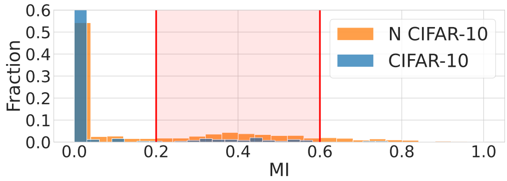


Figure 10: MI of the ensemble for Near OoD (N) CIFAR-10 and real CIFAR-10 samples. We use this to find plausible thresholds of MI.

Model	Test Accuracy	AUROC %			
		F-MNIST (SE)	F-MNIST (SC)	N-MNIST (Ours) (SE)	N-MNIST (Ours) (SC)
LeNet	98.97 ± 0.02	98.87 ± 0.05	98.80 ± 0.05	$\mathbf{65.29 \pm 0.12}$	$\mathbf{64.14 \pm 0.11}$
AlexNet	99.04 ± 0.03	99.10 ± 0.05	99.07 ± 0.05	$\mathbf{70.31 \pm 0.15}$	$\mathbf{69.64 \pm 0.13}$
VGG-11	99.35 ± 0.02	99.20 ± 0.04	99.17 ± 0.03	$\mathbf{72.11 \pm 0.13}$	$\mathbf{71.85 \pm 0.13}$
ResNet-18	99.54 ± 0.02	99.16 ± 0.03	99.14 ± 0.04	$\mathbf{73.15 \pm 0.13}$	$\mathbf{72.81 \pm 0.12}$

Table 4: AUROC % on MNIST using softmax entropy (SE) and softmax confidence (SC) with Fashion(F)-MNIST and Near OoD(N)-MNIST as OoD.

Model	Test/Val Set Accuracy		
	CIFAR-10	CIFAR-100	ImageNet
DenseNet-121	95.66 ± 0.05	80.08 ± 0.15	–
ResNet-50	95.34 ± 0.05	78.26 ± 0.33	80.86
ResNet-110	95.50 ± 0.12	79.50 ± 0.27	–
VGG-16	93.81 ± 0.09	74.33 ± 0.18	–
Wide-ResNet-28-10	96.33 ± 0.07	80.60 ± 0.11	–
Wide-ResNet-50-2	–	–	81.60
Inception-v3	95.25 ± 0.10	78.04 ± 0.14	–
ViT-B-16	99.12 ± 0.03	92.73 ± 0.04	85.02
ViT-B-32	98.73 ± 0.01	92.14 ± 0.02	84.72
ViT-L-16	99.18 ± 0.01	93.73 ± 0.04	86.55
ViT-L-32	99.02 ± 0.01	93.29 ± 0.04	85.47

Table 5: CIFAR-10/100 test and ImageNet val accuracy for CNNs and ViTs used in our evaluation.

OoD sets. The corresponding AUPRC scores for all models trained on CIFAR-10 and CIFAR-100 are shown in table 12 and table 14 for deterministic models and table 7 and table 9 for deep ensembles respectively. In addition, we also show the AUPRC scores as plots for deterministic models, deep ensembles and Vision Transformers in fig. 11.

Evaluation of Shifted Datasets We present the ECE% of 6 architectures: DenseNet-121, ResNet-50/110, VGG-16, Wide-ResNet-28-10 and Inception-v3, all trained on CIFAR-10 on CIFAR-10-C where we use 15 different corruption types at the highest intensity (i.e., 5) and compare with the ECE% of shifted CIFAR-10. Similarly, we present the ECE% of 4 ViTs: ViT-B-16, ViT-B-32, ViT-L-16 and ViT-L-32, trained on ImageNet, evaluated on ImageNet-C. The results are presented in fig. 12. Clearly, the ECE of shifted CIFAR-10 and shifted ImageNet is significantly higher than all corruption types.

D. Qualitative Examples of Generated Samples

In fig. 13, fig. 14 and fig. 15, we present additional qualitative samples of shifted and near OoD examples respectively for both CIFAR-10, CIFAR-100 and ImageNet. In fig. 13, on the left column, we show real samples from CIFAR-10 and CIFAR-100. On the right column, we show corresponding shifted samples. Similar examples for ImageNet can be found in fig. 14. Finally, in fig. 15, we show examples of near OoD samples for both CIFAR-10, CIFAR-100 and ImageNet.

Model	**AUROC**		
	SVHN	**CIFAR-100**	**N CIFAR-10**
DenseNet-121	97.52	91.42	**85.67**
ResNet-50	96.24	90.89	**84.66**
ResNet-110	96.75	91.3	**85.55**
VGG-16	91.26	89.16	**81.07**
Wide-ResNet-28-10	96.59	91.78	**86.54**
Inception-v3	96.12	91.31	**87.07**

Table 6: AUROC of ensemble models trained on CIFAR-10 using predictive entropy on SVHN, CIFAR-100 and Near OoD CIFAR-10 (N CIFAR-10).

Model	**AUPRC**		
	SVHN	**CIFAR-10**	**N CIFAR-100**
DenseNet-121	98.83	90.72	**86.92**
ResNet-50	97.87	89.7	**85.83**
ResNet-110	98.19	90.35	**86.7**
VGG-16	94.89	87.97	**83.93**
Wide-ResNet-28-10	98.06	90.94	**88.03**
Inception-v3	97.79	90.14	**88.14**

Table 7: AUPRC of ensemble models trained on CIFAR-10 using predictive entropy on SVHN, CIFAR-100 and Near OoD CIFAR-10 (N CIFAR-10).

Model	**AUROC**		
	SVHN	**CIFAR-10**	**N CIFAR-100**
DenseNet-121	88.75	82.92	**62.33**
ResNet-50	82.66	81.28	**57.18**
ResNet-110	81.07	82.55	**59.7**
VGG-16	78.3	78.87	**52.42**
Wide-ResNet-28-10	83.62	82.65	**62.83**
Inception-v3	83.89	83.3	**64.66**

Table 8: AUROC of ensemble models trained on CIFAR-100 using predictive entropy on SVHN, CIFAR-10 and Near OoD CIFAR-100.

Model	AUPRC		
	SVHN	**CIFAR-10**	**N CIFAR-100**
DenseNet-121	93.97	78.96	**67.82**
ResNet-50	90.24	76.78	**64.85**
ResNet-110	89.08	78.67	**66.21**
VGG-16	88.27	74.51	**62.7**
Wide-ResNet-28-10	91.27	78.81	**67.94**
Inception-v3	89.81	79.14	**68.76**

Table 9: AUPRC of ensemble models trained on CIFAR-100 using predictive entropy on SVHN, CIFAR-10 and Near OoD CIFAR-100.

Model	**CIFAR-10-C**		**Shifted CIFAR-10 (Ours)**
	Avg ECE %	***Max ECE %***	***ECE %***
DenseNet121	13.69 ± 0.17	25.86 ± 0.40	$\mathbf{51.55 \pm 0.33}$
ResNet-50	13.71 ± 0.48	25.76 ± 1.09	$\mathbf{50.07 \pm 1.24}$
ResNet-110	14.40 ± 0.28	28.03 ± 0.55	$\mathbf{52.16 \pm 0.66}$
VGG-16	17.51 ± 0.22	34.45 ± 0.40	$\mathbf{56.25 \pm 0.41}$
Wide-ResNet-28-10	11.92 ± 0.13	22.87 ± 0.21	$\mathbf{49.64 \pm 0.43}$
Inception-v3	13.47 ± 0.37	25.10 ± 0.71	$\mathbf{52.84 \pm 0.19}$

Table 10: ECE % on CIFAR-10-C compared to Shifted CIFAR-10.

Model	**AUROC SVHN**			**AUROC CIFAR-100**			**AUROC Near OoD CIFAR-10**		
	Entropy	***Confidence***	***Mahalanobis***	***Entropy***	***Confidence***	***Mahalanobis***	***Entropy***	***Confidence***	***Mahalanobis***
DenseNet-121	93.12 ± 1.13	92.85 ± 1.11	96.22 ± 0.30	87.23 ± 0.21	87.17 ± 0.22	89.71 ± 0.14	78.81 ± 0.36	79.11 ± 0.34	79.75 ± 0.39
ResNet-50	92.39 ± 0.30	92.17 ± 0.30	92.67 ± 1.35	86.92 ± 0.53	86.78 ± 0.50	88.40 ± 0.33	78.92 ± 0.75	79.09 ± 0.72	79.04 ± 0.57
ResNet-110	91.63 ± 1.82	91.41 ± 1.81	91.94 ± 1.56	87.48 ± 0.09	87.35 ± 0.09	87.91 ± 0.2	78.08 ± 0.49	80.20 ± 0.48	78.14 ± 0.50
VGG-16	86.70 ± 1.05	86.78 ± 1.00	90.93 ± 0.81	83.37 ± 0.22	83.30 ± 0.21	85.94 ± 0.35	73.43 ± 0.55	73.61 ± 0.53	75.46 ± 1.12
Wide-ResNet-28-10	90.98 ± 1.14	90.89 ± 1.09	98.72 ± 0.11	88.60 ± 0.06	88.48 ± 0.06	91.15 ± 0.02	80.56 ± 0.47	81.73 ± 0.46	81.78 ± 0.11
Inception-v3	91.94 ± 0.54	91.77 ± 0.53	93.49 ± 0.79	86.54 ± 0.43	86.42 ± 0.42	89.56 ± 0.28	80.27 ± 0.39	80.41 ± 0.38	83.76 ± 0.43
ViT-B-16	99.65 ± 0.01	99.49 ± 0.01	96.67 ± 0.18	98.33 ± 0.03	98.19 ± 0.03	98.87 ± 0.00	87.00 ± 0.04	$87.08 \pm .04$	86.65 ± 0.22
ViT-B-32	99.65 ± 0.01	99.44 ± 0.02	95.35 ± 0.21	98.10 ± 0.03	97.93 ± 0.03	98.67 ± 0.01	85.33 ± 0.12	$85.44 \pm .12$	86.21 ± 0.23
ViT-L-16	99.76 ± 0.02	99.64 ± 0.01	$97.66 \pm .42$	98.70 ± 0.02	98.61 ± 0.01	99.17 ± 0.01	85.93 ± 0.28	86.15 ± 0.27	89.47 ± 0.25
ViT-L-32	99.78 ± 0.01	99.63 ± 0.02	95.63 ± 0.09	$98.45 \pm .02$	$98.29 \pm .02$	$98.80 \pm .02$	85.25 ± 0.2	85.38 ± 0.20	84.83 ± 0.11

Table 11: AUROC of models trained on CIFAR-10 using softmax entropy (Entropy), softmax confidence (Confidence) and Mahalanobis distance on SVHN, CIFAR-100 and Near OoD CIFAR-10. Near OoD samples are far harder to detect given their consistently low AUROC scores.

Model	**AUPRC SVHN**			**AUPRC CIFAR-100**			**AUPRC Near OoD CIFAR-10**		
	Entropy	***Confidence***	***Mahalanobis***	***Entropy***	***Confidence***	***Mahalanobis***	***Entropy***	***Confidence***	***Mahalanobis***
DenseNet-121	96.78 ± 0.38	82.89 ± 5.11	94.4 ± 0.41	86.84 ± 0.11	84.7 ± 0.66	89.75 ± 0.15	80.38 ± 0.14	63.6 ± 0.52	68.72 ± 0.75
ResNet-50	95.88 ± 0.13	86.19 ± 1.44	88.41 ± 2.29	85.85 ± 0.39	85.39 ± 0.92	88.45 ± 0.42	80.97 ± 0.33	65.81 ± 1.92	68.86 ± 1.09
ResNet-110	95.58 ± 0.95	85.35 ± 3.25	86.25 ± 2.29	86.29 ± 0.14	86.27 ± 0.23	87.54 ± 0.29	80.8 ± 0.28	67.73 ± 1.26	64.41 ± 1.2
Wide-ResNet-28-10	95.58 ± 0.59	77.81 ± 2.88	97.6 ± 0.15	88.15 ± 0.08	85.83 ± 0.16	91.55 ± 0.04	80.9 ± 0.22	66.03 ± 0.87	72.06 ± 0.2
Inception-v3	95.7 ± 0.31	83.32 ± 1.8	90.91 ± 0.86	85.86 ± 0.41	83.14 ± 0.67	90.13 ± 0.28	81.47 ± 0.22	63.45 ± 0.66	78.11 ± 0.34
ViT-B-16	99.86 ± 0.0	99.08 ± 0.02	93.2 ± 0.37	98.41 ± 0.03	98.18 ± 0.03	98.77 ± 0.01	92.54 ± 0.03	76.39 ± 0.19	76.72 ± 0.43
ViT-B-32	99.86 ± 0.01	99.0 ± 0.04	90.64 ± 0.42	98.19 ± 0.03	97.88 ± 0.02	98.57 ± 0.01	91.53 ± 0.08	73.21 ± 0.15	76.56 ± 0.6
ViT-L-16	99.9 ± 0.01	99.36 ± 0.04	95.42 ± 0.85	98.8 ± 0.01	98.49 ± 0.02	99.02 ± 0.01	92.68 ± 0.12	70.26 ± 0.77	81.19 ± 0.29
ViT-L-32	99.91 ± 0.0	99.33 ± 0.02	90.36 ± 0.18	98.55 ± 0.02	98.25 ± 0.04	98.6 ± 0.03	92.02 ± 0.08	69.81 ± 0.59	73.29 ± 0.14

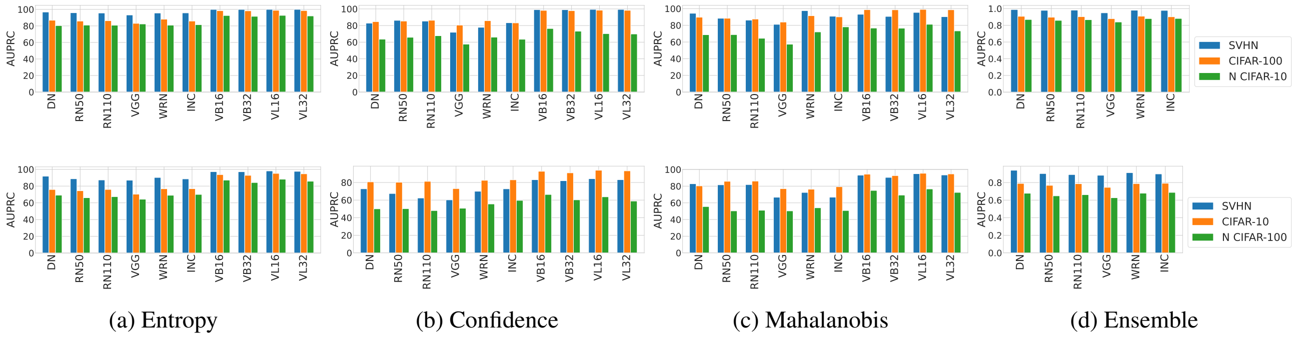
Table 12: AUPRC of models trained on CIFAR-10 using softmax entropy (Entropy), softmax confidence (Confidence) and Mahalanobis distance on SVHN, CIFAR-100 and Near OoD CIFAR-10. Near OoD samples are far harder to detect given their consistently low AUPRC scores.

Model	AUROC SVHN			AUROC CIFAR-100			AUROC Near OoD CIFAR-100		
	Entropy	*Confidence*	*Mahalanobis*	*Entropy*	*Confidence*	*Mahalanobis*	*Entropy*	*Confidence*	*Mahalanobis*
DenseNet-121	84.52 ± 1.55	83.13 ± 1.44	89.49 ± 0.55	79.73 ± 0.25	79.12 ± 0.24	77.26 ± 0.47	60.32 ± 0.27	60.5 ± 0.26	64.16 ± 0.28
ResNet-50	79.77 ± 0.69	78.82 ± 0.71	86.41 ± 0.11	78.82 ± 0.08	78.26 ± 0.09	82.68 ± 0.18	56.58 ± 0.26	56.67 ± 0.26	58.48 ± 0.67
ResNet-110	77.84 ± 1.56	77.26 ± 1.4	86.62 ± 0.23	79.92 ± 0.17	79.3 ± 0.15	82.9 ± 0.23	58.6 ± 0.56	58.58 ± 0.46	59.73 ± 0.59
VGG-16	76.33 ± 1.12	75.38 ± 0.97	78.01 ± 1.24	74.02 ± 0.14	73.62 ± 0.13	74.99 ± 0.13	51.06 ± 0.14	51.53 ± 0.15	56.41 ± 0.42
Wide-ResNet-28-10	81.85 ± 0.79	80.71 ± 0.7	84.18 ± 1.01	80.82 ± 0.11	80.41 ± 0.12	73.42 ± 0.14	62.19 ± 0.17	62.05 ± 0.14	62.38 ± 0.1
Inception-v3	81.6 ± 1.64	80.95 ± 1.46	81.8 ± 0.57	81.24 ± 0.18	80.89 ± 0.18	79.87 ± 0.22	63.96 ± 0.85	63.39 ± 0.78	60.53 ± 0.97
ViT-B-16	93.31 ± 0.21	91.92 ± 0.19	95.91 ± 0.03	93.29 ± 0.04	92.35 ± 0.05	93.95 ± 0.03	79.47 ± 0.06	79.04 ± 0.06	82.91 ± 0.07
ViT-B-32	92.98 ± 0.13	91.56 ± 0.11	93.78 ± 0.21	91.97 ± 0.2	90.94 ± 0.21	92.22 ± 0.19	75.36 ± 0.16	75.05 ± 0.15	78.97 ± 0.24
ViT-L-16	95.11 ± 0.16	94.29 ± 0.15	97.6 ± 0.04	94.62 ± 0.08	94.04 ± 0.09	95.31 ± 0.09	80.36 ± 0.08	80.23 ± 0.1	84.72 ± 0.21
ViT-L-32	94.01 ± 0.07	92.62 ± 0.06	96.01 ± 0.12	94.09 ± 0.07	93.28 ± 0.06	94.15 ± 0.06	76.87 ± 0.12	76.64 ± 0.12	81.29 ± 0.14

Table 13: AUROC of models trained on CIFAR-100 using softmax entropy (Entropy), softmax confidence (Confidence) and Mahalanobis distance on SVHN, CIFAR-10 and Near OoD CIFAR-100. Near OoD samples are far harder to detect given their consistently low AUROC scores.

Model	AUPRC SVHN			AUPRC CIFAR-100			AUPRC Near OoD CIFAR-100		
	Entropy	*Confidence*	*Mahalanobis*	*Entropy*	*Confidence*	*Mahalanobis*	*Entropy*	*Confidence*	*Mahalanobis*
DenseNet-121	91.84 ± 0.93	72.78 ± 2.29	82.85 ± 0.73	75.89 ± 0.3	80.67 ± 0.38	80.31 ± 0.26	69.17 ± 0.13	50.01 ± 0.6	55.53 ± 0.3
ResNet-50	88.69 ± 0.28	67.45 ± 1.23	81.58 ± 0.51	74.4 ± 0.14	80.26 ± 0.18	85.78 ± 0.45	66.07 ± 0.15	50.27 ± 0.6	50.33 ± 0.91
ResNet-110	87.29 ± 1.01	62.34 ± 2.97	81.9 ± 0.78	75.94 ± 0.17	81.25 ± 0.23	86.1 ± 0.55	67.37 ± 0.35	48.07 ± 1.02	51.08 ± 0.99
VGG-16	87.04 ± 0.78	60.21 ± 1.61	66.7 ± 1.54	70.35 ± 0.18	73.07 ± 0.23	77.02 ± 0.25	64.3 ± 0.07	50.72 ± 0.17	50.32 ± 0.39
Wide-ResNet-28-10	90.27 ± 0.55	69.98 ± 1.23	72.42 ± 0.92	76.76 ± 0.17	82.41 ± 0.11	76.36 ± 0.15	68.98 ± 0.09	55.56 ± 0.19	54.03 ± 0.13
Inception-v3	88.54 ± 1.19	72.75 ± 2.25	66.81 ± 2.55	76.89 ± 0.3	82.99 ± 0.1	79.31 ± 0.35	69.98 ± 0.34	59.6 ± 1.21	50.77 ± 1.07
ViT-B-16	97.24 ± 0.08	83.16 ± 0.85	93.22 ± 0.06	93.73 ± 0.03	92.68 ± 0.06	94.4 ± 0.02	87.15 ± 0.02	66.33 ± 0.15	74.84 ± 0.18
ViT-B-32	97.1 ± 0.06	82.0 ± 0.69	90.46 ± 0.28	92.79 ± 0.18	90.93 ± 0.21	92.66 ± 0.25	84.23 ± 0.15	60.24 ± 0.12	69.31 ± 0.4
ViT-L-16	98.08 ± 0.06	84.33 ± 0.89	94.92 ± 0.09	95.11 ± 0.06	93.89 ± 0.12	95.43 ± 0.08	88.25 ± 0.05	63.66 ± 0.21	76.63 ± 0.33
ViT-L-32	97.58 ± 0.01	83.2 ± 0.41	93.4 ± 0.23	94.65 ± 0.06	93.2 ± 0.08	94.66 ± 0.07	85.8 ± 0.06	58.91 ± 0.24	72.46 ± 0.21

Table 14: AUPRC of models trained on CIFAR-100 using softmax entropy (Entropy), softmax confidence (Confidence) and Mahalanobis distance on SVHN, CIFAR-10 and Near OoD CIFAR-100. Near OoD samples are far harder to detect given their consistently low AUPRC scores.



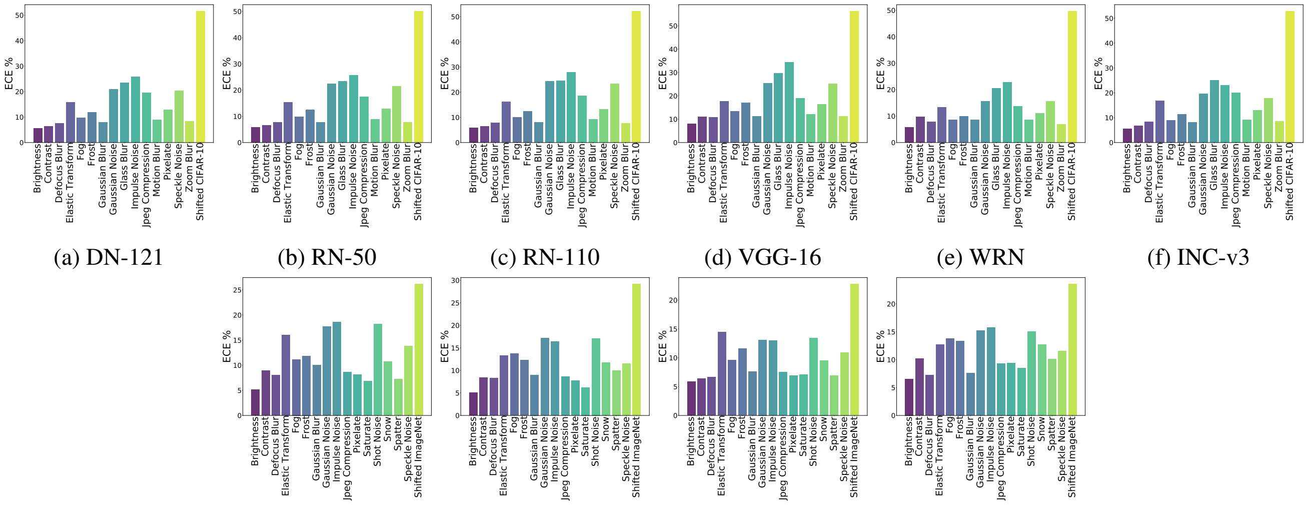
(a) Entropy

(b) Confidence

(c) Mahalanobis

(d) Ensemble

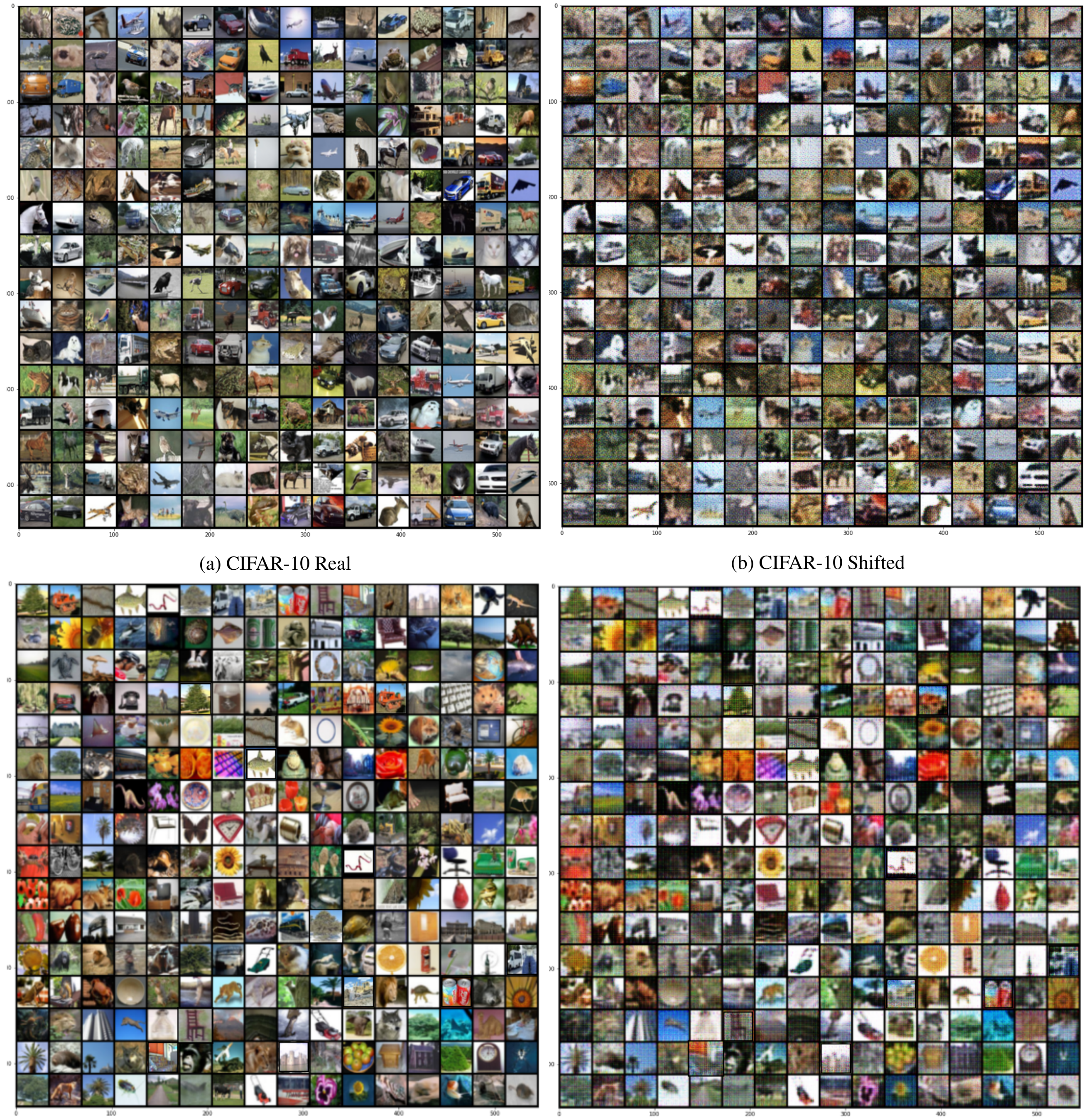
Figure 11: AUPRC % for different models, DenseNet-121 (DN), ResNet-50 (RN50), ResNet-110 (RN110), VGG-16, Wide-ResNet-28-10 (WRN) and Inception-v3 (INC), ViT-B-16/32 (VB16/32) and ViT-L-16/32 (VL16/32) trained on CIFAR-10 (first row) and CIFAR-100 (second row) using SVHN, CIFAR-10/100 and Near OoD (N) CIFAR-10/100 as OoD datasets and softmax entropy, confidence, Mahalanobis distance and deep ensemble baselines.



(a) DN-121 (b) RN-50 (c) RN-110 (d) VGG-16 (e) WRN (f) INC-v3

(g) VB-16 (h) VB-32 (i) VL-16 (j) VL-32

Figure 12: ECE% of models on CIFAR-10-C (top row) and ImageNet-C (bottom row) for different corruption types.



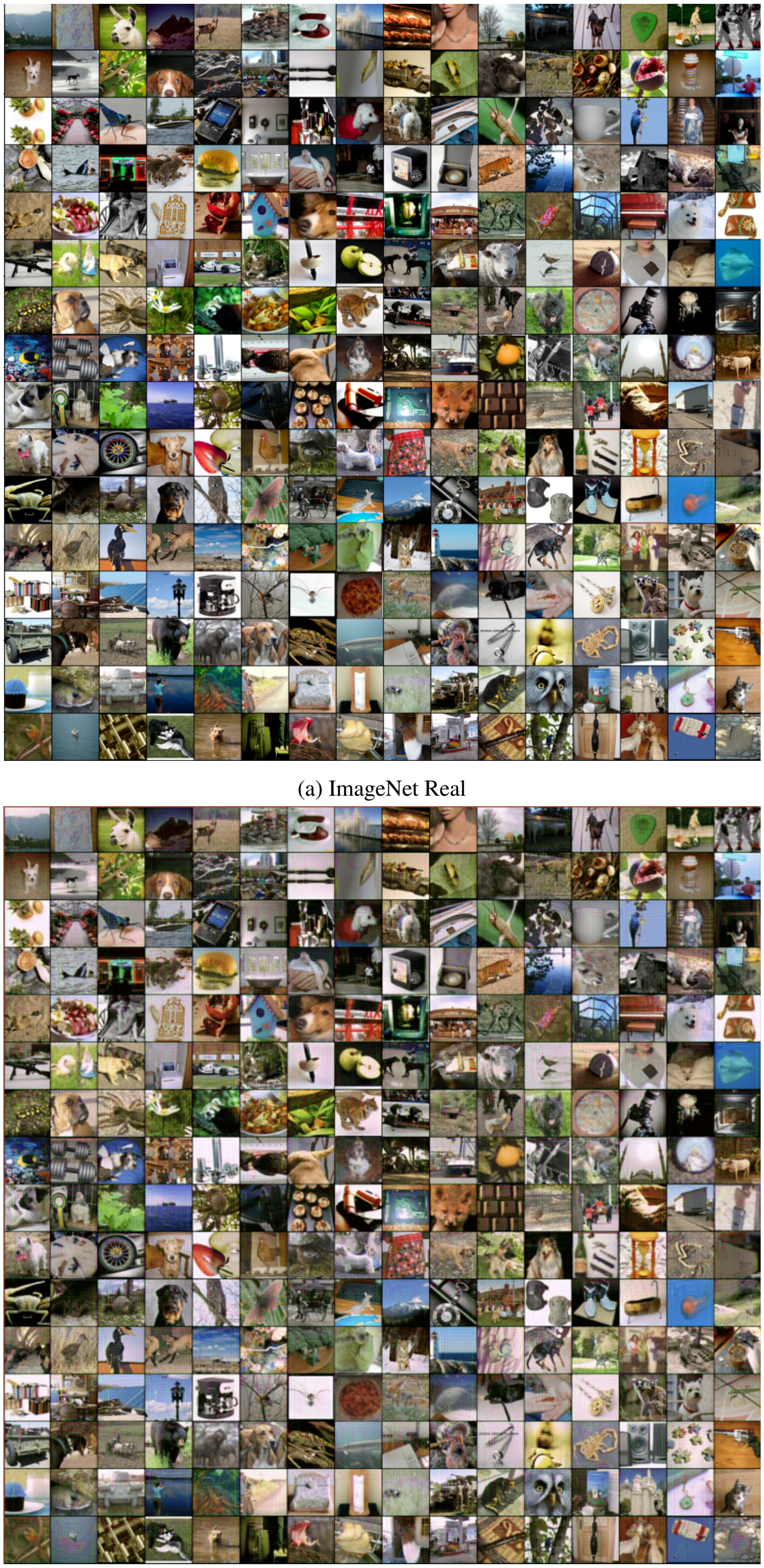
(a) CIFAR-10 Real

(b) CIFAR-10 Shifted

(c) CIFAR-100 Real

(d) CIFAR-100 Shifted

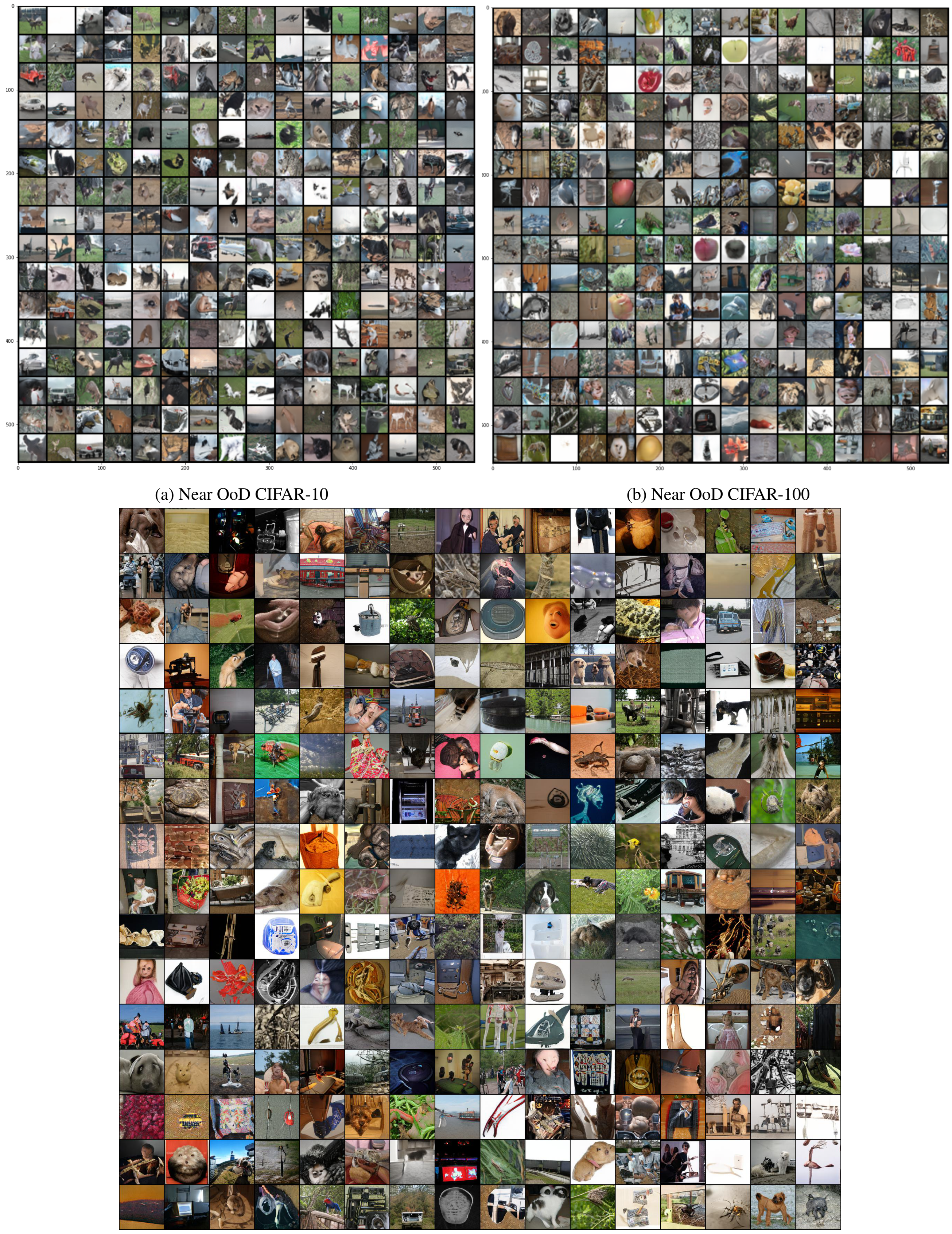
Figure 13: Additional qualitative samples for CIFAR-10 and CIFAR-100 datasets. Left column shows real samples, and the right column shows corresponding shifted/transformed samples.



(a) ImageNet Real

(b) ImageNet Shifted

Figure 14: Additional qualitative samples for ImageNet datasets. Top shows real samples, and bottom shows corresponding shifted/transformed samples.



(a) Near OoD CIFAR-10

(b) Near OoD CIFAR-100

(c) Near OoD ImageNet

Figure 15: Additional qualitative samples for CIFAR-10, CIFAR-100 and ImageNet datasets. Samples show Near OoD images.