Uncertainty-Aware Optimal Transport for Semantically Coherent Out-of-Distribution Detection (Supplementary Materials)

Appendix

A. Experiment Details

The ResNet-18 [1] is employed as the backbone for all experiments, which is trained by an SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9. We use the cosine annealing learning rate starting at 0.1, taking totally 180 epochs. Two dataloaders are prepared with batch-size of 64 and 128 for $\mathcal{D}_L$ and $\mathcal{D}_U$, respectively. For the objective of training is denoted as:

$$L = L_{cls}^{(t)} + \gamma L_{unif}^{(t)} + \lambda L_{rep}, \tag{1}$$

where we set $\gamma = 0.5$ and $\lambda = 0.3$ for all experiments. The number of cluster K for CIFAR-10/100 benchmark is $1024/2048$.

B. Discussion of Training Process

In summary, we alternate the following two steps throughout the training process:

1: Representation learning. Given the updated $D_L^{(t)}$ and $D_U^{(t)}$ based on the assignment matrix $\mathbf{Q}$, the model is trained with Eq. (1) including the inter-cluster extension strategy L_{rep} to obtain a discriminative representation between each ID class and OOD class.

2: Optimizing label assignment. We fix the parameters of the model, and use the model to estimate the energy-based transport cost in the proposed energy-based transport (ET) mechanism. Then we employ the ET to assign correct labels to unlabeled ID samples as many as possible to optimize the assignment matrix $\mathbf{Q}$ with the guidance of the energy-based transport cost.

Notably, the ET is performed at the beginning of the training. Limited by the representation not strong at this stage, the energy metric may not reflect the discrepancy in ID/OOD, thus providing ineffective guidance or even accumulating errors. To explore this doubt, we trained the model only using Eq. (1) for the firstly and performed the above two steps alternately at the remaining epochs, and then evaluated these strategies in Tab. 1. Results show that our method obtains consistently best results across all metrics, which means that performing ET during the whole training process can more fully learn the discrepancy in ID/OOD. In Fig. 1 we also report the comparison among the three experiments above in the number of accurately assigned labels. It can be seen that our method no matter at which epoch can allocate more accurate labels than '50ET' and '150ET', so the ID semantic knowledge hidden in unlabeled set mined by ET at the first few epochs is beneficial to the subsequent training, and the strategy which our method adopts ultimately converges and works.

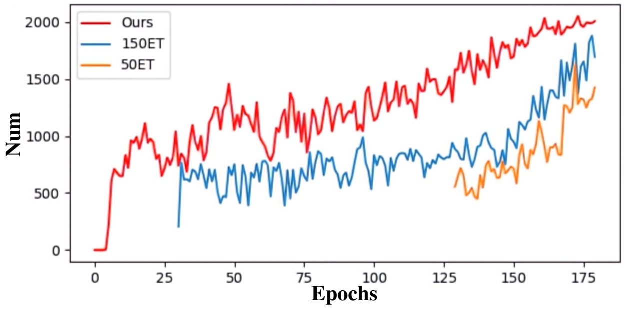


Figure 1. **The number of accurately assigned labels during training in the three experiments in Tab. 1** shows that our method significantly improves over the other two strategies in assigning exact labels to unlabeled ID samples.'50ET' denotes the model performs the two steps mentioned in Appendix B alternately only for the last 50 epochs, while '150ET' alternates the two steps for the last 150 epochs. 'Ours' means our method which alternates the two steps throughout the training process.

C. Effectiveness of the L_{rep}

The inter-cluster extension strategy (L_{rep}) enhances the global feature representation mixed with ID and OOD samples and then the enhanced representation will be mapped into a more discriminate logit space. The energy metrics produced in this space can better reflect the ID/OOD differences to more effectively guide the cluster distribution of ID/OOD samples. In Fig. 2, we use TSNE [8] to visualize the learned feature representation and compare the energy

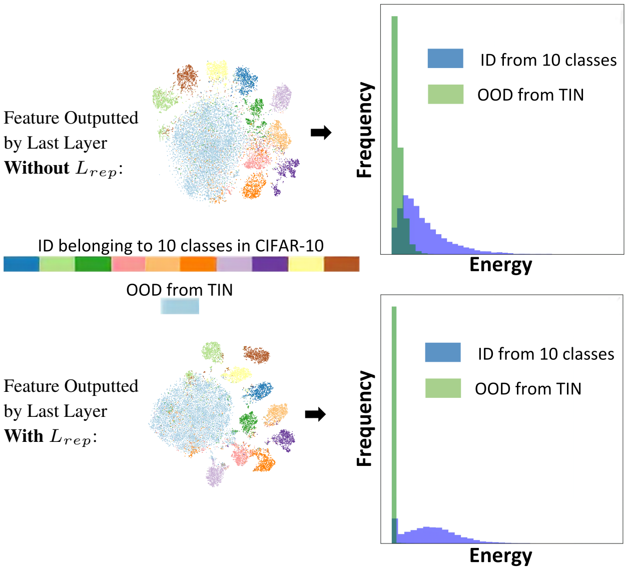


Figure 2. **Comparison of feature representations and the energy metric introduced in the ET between using L_{rep} and without L_{rep}.** L_{rep} not only produces more distinguishable and compact representations, but more importantly, energy metric that reflects ID/OOD differences more significantly can be obtained. From the statistical histogram, it can be seen that when L_{rep} is used, the energy of OOD samples is concentrated at the minimum value and has little overlap with the distribution of ID. These OOD samples will be forced to be uniformly distributed over all clusters in ET. TIN denotes the Tiny-ImageNet dataset.

Table 1. **Comparison between different strategies of training process.** '50ET' denotes the model is trained only with Eq. (1) for the first 130 epochs and performed the two steps mentioned in Appendix B alternately for the last 50 epochs. While '150ET' uses Eq. (1) for training for the first 30 epochs, and performs the two steps alternately for the last 150 epochs. 'Ours' means our method which alternates the two steps throughout the training process. ↑/↓ indicates higher/lower value is better. The best results are in **bold**.

Strategy	FPR95 ↓	AUROC ↑	AUPR-In/Out ↑	ACC ↑
50ET	13.54	93.54	94.72 / 93.53	91.97
150ET	11.83	96.19	95.86 / 94.22	92.87
Ours	**8.53**	**96.47**	**97.10 / 95.65**	**93.71**

metric of ID/OOD in training set with or without L_{rep}. This figure demonstrates the contribution of L_{rep} to the ability of energy metric in the ET to reflect the discrepancy between ID and OOD. Benefiting from the effective guidance of this energy metric for samples with different semantics, L_{rep} ultimately further facilitates ET to explore semantic knowledge hidden in the unlabeled set and improve the performance of the model.

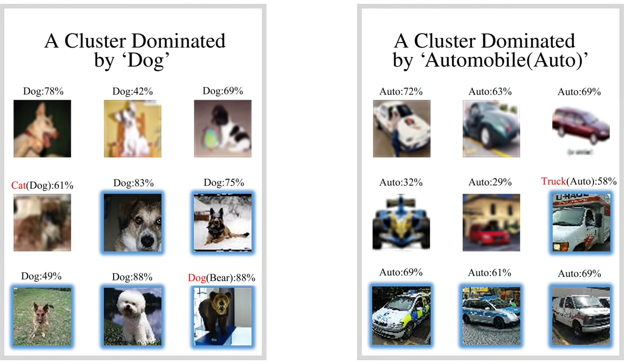


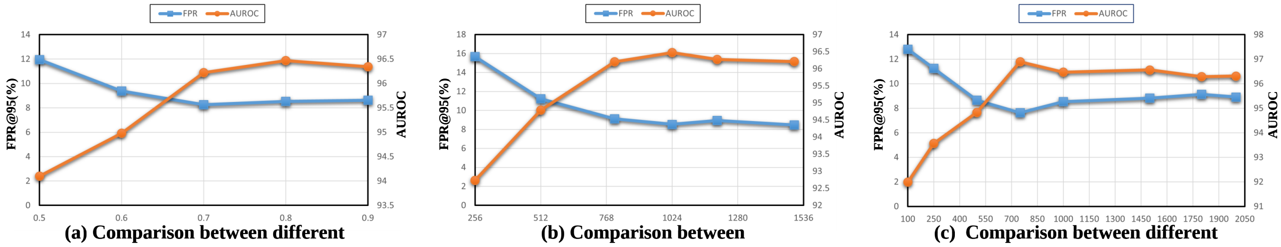
Figure 3. **Visualization of the proposed ET.** We show partial samples from two clusters where the proportion of "Dog" and "Automobile" classes exceed 75%, respectively. The predicted labels and the corresponding prediction probability produced by the model are noted above each image. The red predicted label means that the model classifies the image into a wrong class, and its ground-truth label is in parentheses. The images with blue edges are from the unlabeled Tiny-ImageNet dataset, and the other are from CIFAR-10. The visualization shows that our ET is capable of assigning correct semantic labels to unlabeled ID samples incorrectly predicted by the model or with low confidence.

D. Visualization of the ET

Considering the overconfident prediction of deep neural network models on OOD inputs revealed in [6,7], we cannot assign labels to unlabeled samples relying on the predicted results of the network. Moreover, in the experiment we found that the network will predict the ID samples into wrong classes or output insignificant confidence (softmax probability) on the correct classes. We demonstrate the superiority of the proposed ET in assigning accurate labels through the visualization in Fig. 3, it can be seen from where that our ET can collect unlabeled ID images in a correct manner. This strategy splits the ID samples incorrectly predicted by the model (refer to the two images being predicted into 'cat' and 'truck') or with low confidence (around 30%) from the unlabeled set, and allocate accurate labels to them. It is also noted that a few OOD samples (such as the 'bear' image in Fig. 3 being predicted into 'dog' with overconfidence) are mixed, but it will be corrected at next epochs. To sum up, the proposed ET dramatically improves the reliability of the model in OOD detection tasks, and finally makes our method converge.

E. Choice of hyper-parameters.

Here we analyze the impact of the main hyper-parameters including the threshold of class proportion τ, clusters numbers K, and temperature value T, and prove the robustness of the proposed method. Fig. 4a



(a) Comparison between different thresholds of class proportions

(b) Comparison between different Cluster numbers

(c) Comparison between different temperature values in T-energy

Figure 4. **Discuss about hyper-parameters.** (a), (b), and (c) respectively demonstrate the impact of the threshold of class proportion τ, clusters numbers K, and temperature value T on the performance of our method. All the performance fluctuation is very small when $\tau \geq 0.7$, $K \geq 800$ and $T \geq 750$, showing the robustness of our method.

Table 2. **Comparison between the previous SOTA methods and ours on a large-scale benchmark.** Our method obtains the best results across almost all OOD detection metrics. ↑/↓ indicates higher/lower value is better and the best results are in **bold**.

ID data	Method	FPR95 ↓	AUROC ↑	AUPR-In/Out ↑	CCR@FPR ↑			
					10^{-4}	10^{-3}	10^{-2}	10^{-1}
100 classes from ImageNet	ODIN [4]	80.07	51.23	56.93 / 50.46	0.06	1.18	8.42	13.36
	EBO [5]	82.77	50.17	55.31 / 49.84	0.49	1.49	8.87	13.59
	OE [3]	80.52	55.87	55.93 / 51.94	1.06	**2.63**	7.67	15.13
	MCD [10]	91.04	52.26	54.80 / 43.92	0.08	1.92	5.57	14.35
	UDG [9]	81.89	54.74	57.85 / 52.53	0.95	2.06	9.18	16.35
	Ours	**60.17**	**63.91**	**69.55 / 58.23**	**1.08**	2.16	**10.56**	**21.34**

shows that the maximum fluctuation is 0.38%/0.25% on FPR@95/AUROC when τ=0.7-0.9, and Fig. 4b indicates when K=800-1200, the maximum fluctuation is 0.58%/0.27% on FPR@95/AUROC. We choose τ=0.8 and K=1024 following [9] in the paper for fair comparisons. During rebuttal, we added the experiment shown in Fig. 4c and demonstrate the results are insensitive to large temperature values T (*e.g.,* $\geq$ 750). The best T is around 750 but we also achieve good results when choosing T=1000 in the paper.

F. Experiments on large-scale datasets.

To evaluate the generalization of our method in realistic scenarios, we extend it to large-scale datasets. Specially, we choose 50,000 samples from 100 classes in ImageNet as labeled ID training set D_L and still use Tiny-ImageNet as the unlabeled training set D_U mixed with ID and OOD data. The testing set T contains 1,000 ID images and 4,000 OOD images from ImageNet. The comparison results in Tab. 2 demonstrate that our proposed uncertainty-aware optimal transport scheme obtains the best results on almost all metrics, indicating its generalization on large-scale datasets. However, all methods in Tab. 2 exhibit significant performance degradation on the large-scale datasets, which suggests that more effort is needed to address the challenges presented by benchmarks with more diverse categories and higher resolution in the OOD detection area.

G. Detailed Results and More Architectures

Tab. 3 and Tab. 5 show the detailed results among all datasets. Our method obtains consistently better results across all OOD detection metrics and all datasets. Compared with other methods using extra OOD training data (MCD [10], OE [3], UDG [9]), our method boosts the OOD detection performance meanwhile maximally maintaining the ID classification performance and achieves the best results on ACC. Following [9] we also adopt another network architecture of WideResNet-28 [2] to do experiments in Tab. 4 and Tab. 6 and compare the performance. The comparison results on WideResNet-28 have the same trend as that on ResNet-18 [1] architecture. Our proposed method combining ET with L_{rep} has advantages on almost all metrics, showing that our method enhances both the OOD detection and the ID classification ability. Notably, the previous state-of-the-art approaches generally performed well on SVHN and Texture datasets, but in Tiny-ImageNet, LSUN, and Places365 suffered a defeat. It can be explained that the images in the first two datasets have relatively flat backgrounds, which are quite different in style from those in CIFAR10/100, and the resulting covariate shifts make it easier for the model to identify them as OOD examples. Our proposed method especially achieves performance ad-

Table 3. **Detailed results on CIFAR-10 benchmark using ResNet-18.** Our method obtains consistently better results across almost all OOD detection metrics and all datasets. ACC shows the classification accuracy on all the ID test samples from T^I. ↑/↓ indicates higher/lower value is better.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				ACC ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
ODIN	Texture	42.52	84.06	86.01 / 80.73	0.02	0.18	3.71	40.14	95.02
	SVHN	52.27	83.26	63.76 / 92.60	1.01	4.00	11.82	44.85	95.02
	CIFAR-100	56.34	78.40	73.21 / 80.99	0.10	0.38	4.43	30.11	95.02
	Tiny-ImageNet	59.09	79.69	79.34 / 77.52	0.36	0.63	4.49	34.52	92.54
	LSUN	47.85	84.56	81.56 / 85.58	0.21	0.85	9.92	46.95	95.02
	Places365	53.94	82.01	54.92 / 93.30	0.47	1.68	7.13	39.63	93.87
	Mean	**52.00**	**82.00**	**73.13 / 85.12**	**0.36**	**1.29**	**6.92**	**39.37**	**94.42**
EBO	Texture	52.11	80.70	83.34 / 75.20	0.01	0.13	2.79	31.96	95.02
	SVHN	30.56	92.08	80.95 / 96.28	1.85	5.74	21.44	75.81	95.02
	CIFAR-100	56.98	79.65	75.09 / 81.23	0.10	0.69	4.74	34.28	95.02
	Tiny-ImageNet	57.81	81.65	81.80 / 78.75	0.33	0.95	6.01	40.40	92.54
	LSUN	50.56	85.04	82.80 / 85.29	0.24	1.96	11.35	50.43	95.02
	Places365	52.16	83.86	58.96 / 93.90	0.39	2.11	8.38	46.00	93.87
	Mean	**50.03**	**83.83**	**77.15 / 85.11**	**0.49**	**1.93**	**9.12**	**46.48**	**94.42**
MCD	Texture	83.92	81.59	90.20 / 63.27	4.97	10.51	29.52	62.10	90.56
	SVHN	60.27	89.78	85.33 / 94.25	20.05	38.23	55.43	74.01	90.56
	CIFAR-100	74.00	82.78	83.97 / 79.16	0.80	4.99	18.88	58.18	90.56
	Tiny-ImageNet	78.89	80.98	85.63 / 72.48	1.62	4.15	19.37	56.08	87.33
	LSUN	68.96	84.71	85.74 / 81.50	1.75	7.93	21.88	61.54	90.56
	Places365	72.08	83.51	69.44 / 92.52	3.29	7.97	23.07	60.22	88.51
	Mean	**73.02**	**83.89**	**83.39 / 80.53**	**5.41**	**12.30**	**28.02**	**62.02**	**89.68**
OE	Texture	51.17	89.56	93.79 / 81.88	6.58	11.80	27.99	71.13	91.87
	SVHN	20.88	96.43	93.62 / 98.32	32.72	47.33	67.20	86.75	91.87
	CIFAR-100	58.54	86.22	86.17 / 84.88	3.64	6.55	19.04	61.11	91.87
	Tiny-ImageNet	58.98	87.65	90.9 / 82.16	14.37	18.84	33.65	66.03	89.27
	LSUN	57.97	86.75	87.69 / 85.07	11.8	19.62	29.22	61.95	91.87
	Places365	55.64	87.00	73.11 / 94.67	11.36	17.36	26.33	62.23	90.99
	Mean	**50.53**	**88.93**	**87.55 / 87.83**	**13.41**	**20.25**	**33.91**	**68.20**	**91.29**
UDG	Texture	20.43	96.44	98.12 / 92.91	19.90	43.33	69.19	87.71	92.94
	SVHN	13.26	97.49	95.66 / 98.69	36.64	56.81	76.77	89.54	92.94
	CIFAR-100	47.20	90.98	91.74 / 89.36	1.50	10.94	40.34	75.89	92.94
	Tiny-ImageNet	50.18	91.91	94.43 / 86.99	0.32	23.15	53.96	78.36	90.22
	LSUN	42.05	93.21	94.53 / 91.03	14.26	37.59	60.62	81.69	92.94
	Places365	44.22	92.64	87.17 / 96.66	10.62	35.05	58.96	79.63	91.68
	Mean	**36.22**	**93.78**	**93.61 / 92.61**	**13.87**	**34.48**	**59.97**	**82.14**	**92.28**
Ours	Texture	0.78	98.92	99.55 / 97.73	49.52	72.02	87.65	91.10	94.66
	SVHN	0.26	99.03	98.79 / 99.91	64.87	82.96	91.10	93.44	94.66
	CIFAR-100	29.17	91.17	92.04 / 90.02	3.16	14.07	34.89	72.02	94.66
	Tiny-ImageNet	7.15	94.15	97.17 / 88.57	64.27	78.25	81.40	84.20	90.41
	LSUN	0.53	98.93	99.28 / 98.91	34.88	77.74	87.65	91.10	94.66
	Places365	13.26	96.61	95.71 / 98.78	25.16	58.64	81.40	85.79	93.23
	Mean	**8.53**	**96.47**	**97.10 / 95.65**	**40.31**	**63.95**	**77.35**	**86.27**	**93.71**

Table 4. **Detailed results on CIFAR-10 benchmark using WideResNet-28.** Our method obtains consistently better results across almost all OOD detection metrics and all datasets. ACC shows the classification accuracy on all the ID test samples from T^I. ↑/↓ indicates higher/lower value is better.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				ACC ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
ODIN	Texture	47.50	81.23	82.94 / 78.25	0.00	0.00	1.81	32.69	96.08
	SVHN	51.17	85.36	68.02 / 93.53	1.10	3.54	13.08	53.04	96.08
	CIFAR-100	52.92	79.47	73.57 / 82.59	0.00	0.36	3.97	30.55	96.08
	Tiny-ImageNet	54.86	80.39	78.82 / 79.48	0.01	0.36	3.12	33.69	93.69
	LSUN	46.53	81.86	75.70 / 85.03	0.25	0.68	3.91	33.49	96.08
	Places365	49.03	81.49	49.84 / 93.60	0.04	0.55	3.72	33.14	95.02
	Mean	**50.33**	**81.63**	**71.48 / 85.41**	**0.23**	**0.91**	**4.94**	**36.10**	**95.51**
EBO	Texture	40.44	89.55	91.16 / 84.41	0.00	0.00	5.41	71.35	96.08
	SVHN	16.13	96.90	93.77 / 98.47	2.93	18.26	68.48	91.28	96.08
	CIFAR-100	42.41	88.97	85.73 / 89.42	0.01	0.72	8.77	67.94	96.08
	Tiny-ImageNet	45.81	89.55	89.55 / 86.72	0.03	0.61	9.93	73.79	93.69
	LSUN	37.14	90.58	87.47 / 91.07	0.29	0.83	8.51	76.21	96.08
	Places365	39.84	89.86	68.32 / 96.33	0.04	0.68	7.15	73.24	95.02
	Mean	**36.96**	**90.90**	**86.00 / 91.07**	**0.55**	**3.52**	**18.04**	**75.64**	**95.51**
MCD	Texture	93.19	70.58	82.49 / 49.12	0.00	0.15	7.65	44.96	87.85
	SVHN	88.68	81.37	74.43 / 86.75	3.28	8.65	28.28	66.86	87.85
	CIFAR-100	83.29	76.58	77.17 / 72.50	0.03	0.72	10.47	45.36	87.85
	Tiny-ImageNet	86.6	74.83	80.53 / 64.30	0.04	2.48	12.88	44.47	85.58
	LSUN	93.06	70.14	72.62 / 63.38	0.55	2.81	10.51	36.16	87.85
	Places365	93.13	70.42	49.04 / 84.32	0.10	2.39	9.65	36.37	86.48
	Mean	**89.66**	**73.99**	**72.71 / 70.06**	**0.67**	**2.87**	**13.24**	**45.7**	**87.24**
OE	Texture	35.14	92.44	95.27 / 87.17	5.27	8.94	31.17	79.23	94.95
	SVHN	22.94	96.23	94.14 / 97.78	37.34	52.79	73.87	88.74	94.95
	CIFAR-100	52.99	87.17	86.80 / 86.09	1.72	6.83	21.22	63.16	94.95
	Tiny-ImageNet	55.53	87.43	90.20 / 82.58	4.58	13.91	28.61	64.92	92.72
	LSUN	59.69	85.56	86.18 / 83.67	5.18	11.55	26.09	58.88	94.95
	Places365	55.30	85.75	69.15 / 94.25	4.50	10.31	22.42	56.79	94.24
	Mean	**46.93**	**89.10**	**86.96 / 88.59**	**9.76**	**17.39**	**33.90**	**68.62**	**94.46**
UDG	Texture	22.59	95.86	97.49 / 92.59	0.87	8.92	58.06	87.56	94.50
	SVHN	17.23	97.23	95.43 / 98.64	45.32	60.75	78.46	89.84	94.50
	CIFAR-100	43.36	91.53	92.08 / 90.21	5.19	12.28	37.79	77.03	94.50
	Tiny-ImageNet	39.33	93.90	95.90 / 90.01	4.86	27.52	64.17	82.97	92.07
	LSUN	30.17	95.25	96.06 / 94.05	13.28	36.98	66.03	86.35	94.50
	Places365	35.24	94.31	89.24 / 97.55	8.39	27.67	61.10	83.75	93.33
	Mean	**31.32**	**94.68**	**94.36 / 93.84**	**12.98**	**29.02**	**60.93**	**84.58**	**93.90**
Ours	Texture	2.03	99.43	99.65 / 99.02	21.81	71.75	88.68	95.07	95.73
	SVHN	1.13	99.87	99.72 / 99.93	80.10	85.83	94.53	95.61	95.37
	CIFAR-100	31.40	91.43	91.03 / 90.83	9.95	15.73	24.48	77.81	95.73
	Tiny-ImageNet	9.37	97.18	98.35 / 94.21	72.74	80.91	85.43	89.28	92.16
	LSUN	5.18	98.83	98.92 / 98.77	48.62	54.92	82.67	93.73	95.73
	Places365	12.49	97.25	94.15 / 98.92	16.15	46.68	73.43	89.70	94.21
	Mean	**8.24**	**97.33**	**96.97 / 96.95**	**41.56**	**59.26**	**74.87**	**90.20**	**94.88**

Table 5. **Detailed results on CIFAR-100 benchmark using ResNet-18.** Our method obtains consistently better results across almost all OOD detection metrics and all datasets. ACC shows the classification accuracy on all the ID test samples from T^I. ↑/↓ indicates higher/lower value is better.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				ACC ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
ODIN	Texture	79.47	77.92	86.69 / 62.97	2.66	4.66	15.09	45.82	76.65
	SVHN	90.33	75.59	65.25 / 84.49	4.98	12.02	23.79	46.61	76.65
	CIFAR-10	81.82	77.90	79.93 / 73.39	0.09	3.69	15.39	47.20	76.65
	Tiny-ImageNet	82.74	77.58	86.26 / 61.38	0.20	3.78	15.99	45.56	69.56
	LSUN	80.57	78.22	86.34 / 63.44	1.68	5.59	17.37	45.56	76.10
	Places365	76.42	80.66	66.77 / 89.66	1.45	4.16	18.98	49.60	77.56
	Mean	**81.89**	**77.98**	**78.54 / 72.56**	**1.84**	**5.65**	**17.77**	**46.73**	**75.53**
EBO	Texture	84.29	76.32	85.87 / 59.12	0.82	3.89	14.37	44.60	76.65
	SVHN	78.23	83.57	75.61 / 90.24	9.67	17.27	33.70	57.26	76.65
	CIFAR-10	81.25	78.95	80.01 / 74.44	0.05	4.63	18.03	48.67	76.65
	Tiny-ImageNet	83.32	78.34	87.08 / 62.13	1.04	6.37	21.44	47.92	69.56
	LSUN	84.51	77.66	86.42 / 61.40	1.59	6.44	19.58	46.66	76.10
	Places365	78.37	80.99	68.22 / 89.60	1.40	4.94	21.32	51.21	77.56
	Mean	**81.66**	**79.31**	**80.54 / 72.82**	**2.43**	**7.26**	**21.41**	**49.39**	**75.53**
MCD	Texture	83.97	73.46	83.11 / 56.79	0.07	1.03	9.29	38.09	68.80
	SVHN	85.82	76.61	65.50 / 85.52	3.03	8.66	23.15	45.44	68.80
	CIFAR-10	87.74	73.15	76.51 / 67.24	0.35	3.26	16.18	41.41	68.80
	Tiny-ImageNet	84.46	75.32	85.11 / 59.49	0.24	6.14	19.66	41.44	62.21
	LSUN	86.08	74.05	84.21 / 58.62	1.57	5.16	18.05	41.25	67.51
	Places365	82.74	76.30	61.15 / 87.19	1.08	3.35	14.04	43.37	70.47
	Mean	**85.14**	**74.82**	**75.93 / 69.14**	**1.06**	**4.60**	**16.73**	**41.83**	**67.77**
OE	Texture	86.56	73.89	84.48 / 54.84	0.66	2.86	12.86	41.81	70.49
	SVHN	68.87	84.23	75.11 / 91.41	7.33	14.07	31.53	54.62	70.49
	CIFAR-10	79.72	78.92	81.95 / 74.28	2.82	9.53	23.90	48.21	70.49
	Tiny-ImageNet	83.41	76.99	86.36 / 60.56	0.22	8.50	21.95	43.98	63.69
	LSUN	83.53	77.10	86.28 / 60.97	1.72	7.91	22.61	44.19	69.89
	Places365	78.24	79.62	67.13 / 88.89	3.69	7.35	20.22	47.68	72.02
	Mean	**80.06**	**78.46**	**80.22 / 71.83**	**2.74**	**8.37**	**22.18**	**46.75**	**69.51**
UDG	Texture	75.04	79.53	87.63 / 65.49	1.97	4.36	9.49	33.84	68.51
	SVHN	60.00	88.25	81.46 / 93.63	14.90	25.50	38.79	56.46	68.51
	CIFAR-10	83.35	76.18	78.92 / 71.15	1.99	5.58	17.27	42.11	68.51
	Tiny-ImageNet	81.73	77.18	86.00 / 61.67	0.67	4.82	17.80	41.72	61.80
	LSUN	78.70	76.79	84.74 / 63.05	1.59	5.34	18.04	44.70	67.10
	Places365	73.86	79.87	65.36 / 89.60	1.96	6.33	22.03	47.97	69.83
	Mean	**75.45**	**79.63**	**80.69 / 74.10**	**3.85**	**8.66**	**20.57**	**44.47**	**67.38**
Ours	Texture	47.85	82.91	90.14 / 67.27	0.56	2.02	23.81	52.10	73.06
	SVHN	7.10	95.43	93.83 / 98.31	25.22	47.87	63.36	68.39	73.06
	CIFAR-10	79.25	68.20	71.40 / 63.78	0.31	3.14	10.75	34.04	73.06
	Tiny-ImageNet	1.95	90.28	95.23 / 77.52	36.64	44.98	55.52	59.88	61.31
	LSUN	54.09	78.34	87.37 / 62.22	4.53	12.77	26.58	44.76	70.69
	Places365	56.08	79.45	68.25 / 89.70	0.60	6.04	24.25	44.62	72.92
	Mean	**41.05**	**82.44**	**84.37 / 76.47**	**11.70**	**19.47**	**34.05**	**50.97**	**70.70**

Table 6. **Detailed results on CIFAR-100 benchmark using WideResNet-28.** Our method obtains consistently better results across almost all OOD detection metrics and all datasets. ACC shows the classification accuracy on all the ID test samples from T^I. ↑/↓ indicates higher/lower value is better.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				ACC ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
ODIN	Texture	78.88	76.46	84.68 / 62.45	0.15	1.52	10.21	41.44	80.25
	SVHN	92.26	68.41	49.07 / 81.28	1.73	2.93	8.02	28.93	80.25
	CIFAR-10	78.22	80.14	81.43 / 76.26	0.06	3.09	15.78	50.75	80.25
	Tiny-ImageNet	80.54	77.88	85.89 / 62.67	0.24	2.25	13.97	45.53	72.92
	LSUN	78.11	78.66	85.57 / 65.68	0.19	1.26	11.69	45.32	78.54
	Places365	73.62	80.57	63.79 / 90.13	0.86	2.79	13.03	47.47	80.03
	Mean	**80.27**	**77.02**	**75.07 / 73.08**	**0.54**	**2.31**	**12.12**	**43.24**	**78.71**
EBO	Texture	84.22	76.13	85.08 / 58.51	0.08	1.55	10.04	44.24	80.25
	SVHN	80.05	79.88	65.44 / 88.37	0.97	3.88	14.93	50.85	80.25
	CIFAR-10	76.18	81.50	83.34 / 77.36	0.45	6.11	21.03	53.73	80.25
	Tiny-ImageNet	80.78	79.94	88.02 / 64.18	0.06	4.92	22.31	51.82	72.92
	LSUN	82.59	78.74	86.71 / 62.94	0.64	1.55	17.71	49.76	78.54
	Places365	74.54	81.63	67.67 / 90.18	1.13	3.69	17.55	52.47	80.03
	Mean	**79.73**	**79.64**	**79.38 / 73.59**	**0.55**	**3.62**	**17.26**	**50.48**	**78.71**
MCD	Texture	91.33	69.03	79.60 / 49.66	0.00	0.29	4.49	32.61	68.80
	SVHN	87.03	73.48	52.89 / 84.73	1.74	2.90	6.68	33.88	68.80
	CIFAR-10	86.89	73.79	76.15 / 68.38	0.26	2.88	13.40	39.94	68.80
	Tiny-ImageNet	85.16	74.59	84.19 / 58.36	1.01	2.58	13.71	40.31	62.22
	LSUN	88.67	72.04	83.06 / 54.33	1.13	3.58	15.95	39.58	67.29
	Places365	86.83	74.05	59.58 / 85.28	1.24	3.66	14.85	41.07	69.77
	Mean	**87.65**	**72.83**	**72.58 / 66.79**	**0.90**	**2.65**	**11.51**	**37.90**	**67.61**
OE	Texture	93.07	67.00	78.92 / 46.52	0.02	0.52	5.50	32.16	74.01
	SVHN	88.74	76.14	66.07 / 85.17	7.06	12.91	24.82	47.43	74.01
	CIFAR-10	78.82	79.36	81.29 / 75.27	1.08	7.63	17.49	48.84	74.01
	Tiny-ImageNet	83.34	78.35	87.34 / 61.78	1.06	8.84	24.40	47.64	66.49
	LSUN	84.96	78.11	87.26 / 60.76	5.80	10.40	25.75	48.27	71.47
	Places365	80.30	79.87	67.23 / 88.65	1.78	6.29	19.78	49.84	74.39
	Mean	**84.87**	**76.47**	**78.02 / 69.69**	**2.80**	**7.76**	**19.63**	**45.70**	**72.40**
UDG	Texture	73.62	79.01	85.53 / 67.08	0.00	0.00	6.74	46.09	75.77
	SVHN	66.76	85.29	76.14 / 92.33	8.00	15.83	32.57	58.05	75.77
	CIFAR-10	82.35	76.67	78.52 / 72.63	0.51	3.90	15.29	44.79	75.77
	Tiny-ImageNet	78.91	79.04	87.00 / 65.06	0.12	2.86	19.13	47.50	68.57
	LSUN	77.04	79.79	87.49 / 66.93	2.51	6.01	22.33	49.14	73.93
	Places365	72.25	81.49	66.72 / 90.65	1.19	3.28	17.59	50.82	76.10
	Mean	**75.16**	**80.21**	**80.23 / 75.78**	**2.05**	**5.31**	**18.94**	**49.40**	**74.32**
Ours	Texture	34.47	73.50	85.15 / 55.77	0.73	4.81	21.51	40.96	76.47
	SVHN	7.71	96.22	93.59 / 97.92	19.61	36.21	58.39	71.37	76.85
	CIFAR-10	56.96	66.97	72.19 / 61.96	2.40	7.66	16.52	33.95	76.82
	Tiny-ImageNet	2.53	89.55	94.95 / 75.10	36.59	42.33	52.94	61.37	66.12
	LSUN	72.63	69.87	81.48 / 52.22	0.67	5.25	14.52	33.03	69.54
	Places365	49.73	79.88	67.82 / 89.04	3.03	8.35	22.07	46.90	76.80
	Mean	**38.19**	**83.01**	**84.90 / 76.60**	**10.20**	**18.57**	**34.17**	**52.70**	**74.89**

vantages on Tiny-ImageNet, LSUN, and Places365, but its performance on CIFAR-100/10 (the styles between them are very similar) of the CIFAR-10/100 benchmark is still not outstanding. This indicates that more exploration is needed to overcome the interference of covariate shifts in OOD detection tasks.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[3] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Int. Conf. Learn. Represent.*, 2019. 3

[4] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Int. Conf. Learn. Represent.*, 2017. 3

[5] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[6] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 427–436, 2015. 2

[7] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. volume 35, pages 1757–1772. IEEE, 2012. 2

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[9] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Int. Conf. Comput. Vis.*, 2021. 3

[10] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Int. Conf. Comput. Vis.*, 2019. 3