A. More Details on SC-OOD Benchmarking

In this section, we will explain more details of the SC-OOD benchmark formation mentioned in Section 4. We first comprehensively describe the difference between the proposed SC-OOD benchmark and DD-OOD benchmark. We scrutinized the existing famous OOD detection benchmarks (referred as DD-OOD) and find that they actually utilize nearest interpolation methods when resizing OOD images into ID image size. As shown in Figure A1, DD-OOD images look more coarse and grainy than ID images, resulting in a detectable sensory difference between 'smooth' ID images and 'coarse' OOD images. In this case, OOD detection methods targeting on DD-OOD benchmark could just impractically focus on low-level covariate shifts and ignore the high-level semantic differences for final decision. Therefore, we aim to propose a more challenging SC-OOD task to actually focus on semantics. In SC-OOD benchmarks, we use the alternative bi-linear interpolation method for resizing, which yields smoother images that are more similar to ID images. We believe it will encourage the models to focus more on semantics for OOD detection, reflecting the purpose of the SC-OOD benchmark. Afterward, we redirect the ID samples from OOD datasets, which has been explained in Section 4.

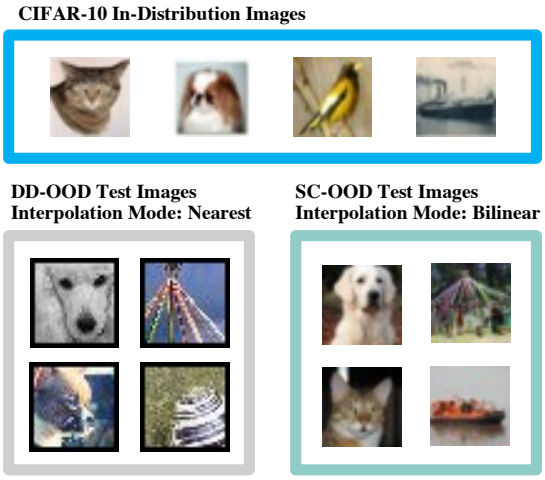

Figure A1: **Exemplar DD-OOD and SC-OOD testing images.** DD-OOD usually utilizes nearest interpolation mode for resizing, which generates grainy images with some sensory differences compared to ID images. SC-OOD takes bi-linear interpolation mode, yielding a more challenging task to encourage SC-OOD methods to focus on semantics.

Table A1: **The record of OOD detection performance as benchmarks gradually changes from DD-OOD to SC-OOD.** It records totally 4 steps from DD-OOD benchmark of CIFAR-10 + Tiny-ImageNet (test, nearest interpolation) to SC-OOD using CIFAR-10 + Tiny-ImageNet (val, bi-linear interpolation) after semantics-based re-splitting.

	FPR95 ↓	AUROC ↑
ODIN	0.46 - 14.3 - 49.9 - 55.0	99.8 - 97.3 - 88.3 - 88.8
EBO	1.56 - 22.8 - 45.6 - 50.6	99.5 - 95.9 - 90.2 - 90.4
MCD	0.01 - 59.1 - 61.5 - 68.6	99.9 - 93.3 - 89.3 - 88.9
UDG	12.3 - 18.3 - 43.7 - 48.3	97.9 - 96.7 - 91.0 - 91.1

Table A1 shows the performance changes from DD-OOD to SC-OOD on CIFAR-10 + Tiny-ImageNet (TIN). Four states are recorded as OOD TIN set gradually changes: **1)** TIN test set, nearest (interpolation), **2)** TIN val set, nearest, **3)** TIN val set, bi-linear, **4)** TIN val set, bi-linear, with re-splitting as SC-OOD eventually. We use TIN val set because it contains ground-truth labels for easier re-splitting. The result shows that even changing test set into validation set will break the perfect performance of some existing methods. Bigger drop exists when interpolation methods change. This drop is understandable since the same interpolation will eliminate all major covariate shifts, but ID and OOD are not yet separated by semantics. However, semantic re-splitting continues to destroy model performance, but UDG gets minimal decrease and better overall scores on both metrics, showing a better understanding of semantics.

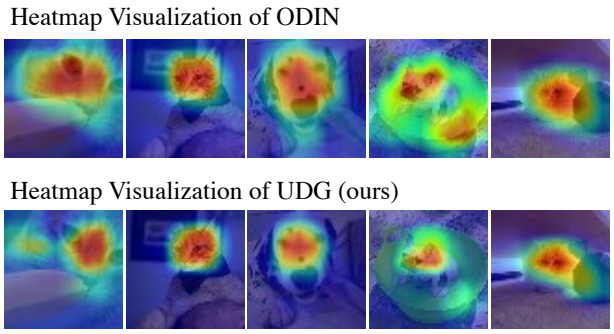

Figure A2: **Heatmap visualization on the images from Figure 3.** The upper part is from ODIN and the lower part is ours. For the fourth image of the dog in the bucket, ODIN is distracted by the irrelevant green bucket for its prediction of dog while ours does not distract. Generally, our method shows better concentration on semantics.

B. Visual Heatmap Comparison

In this section, we visualize the heatmap activated by the previous method ODIN [7] and our proposed UDG on their

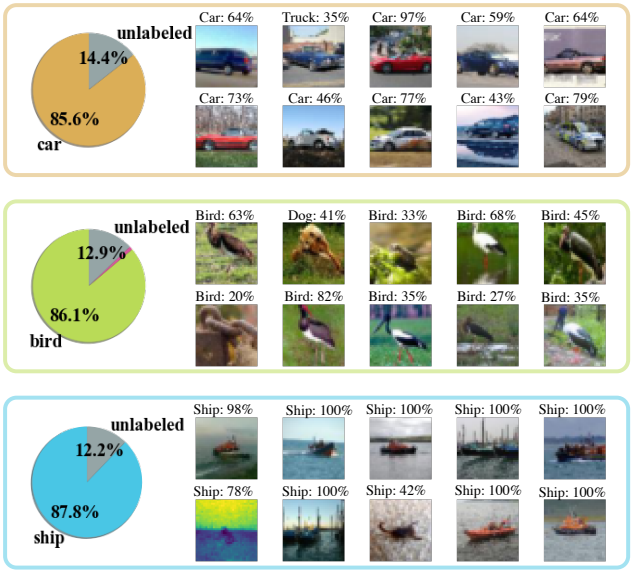

Figure A3: **Visualization of three high group purity clusters for in-distribution filtering (IDF).** We randomly show three clusters with group purity over 0.8 at 80% of the training time. The visualization shows that our IDF strategy accompanied by UDG can filter out ID samples in an accurate manner. The confidence (softmax) score is also presented above each image. Our group-based IDF strategy can also include ID samples with a lower individual confidence score (refer to the last two images of birds).

prediction. We found that the semantic capabilities of UDG are significantly stronger than ODIN, since our model can focus more on the semantic area of the image, while ODIN usually distracts, sometimes even focuses on some irrelevant area (fourth image in Figure A2).

C. Visualization of the Proposed IDF

In this section, we visualize the in-distribution filtering (IDF) process that is described in Section 3.4. According to Figure A3, we find that the major unlabeled data that falls in the high-purity ID group is actually ID samples that belong to the corresponding category. We also notice that this method can also include these images with a relatively low confidence score, for example, for the cluster of birds, the last two images only have confidence around 0.3, which might be difficult to be filtered as ID if we only consider the confidence score. However, our method would be able to include them. According to the second image of cars, even though the network provides an incorrect pseudo label of truck for the car, our IDF strategy can correct the mistake. Even though there are also few mistakes introduced (scorpion in the ship's group), it will be corrected when regrouping in the next epoch. In addition, the overconfidence property of neural networks might give a high confidence score for wrong images, while the filtering strategy of UDG can also help prevent this mistake. In sum, the detailed visualization shows the reliability of the proposed IDF method.

D. Detailed Results and More Architectures

Table A2 and Table A3 expand the average values reported in Table 4. We also do experiments on another network architecture of WideResNet-28 [33]. The result generally has the same trend as ResNet-18 architecture. The proposed UDG method has advantages on almost all the metrics, showing that our method enhances ID classification and OOD detection ability. Notably, the advantages of our proposed method on Tiny-ImageNet, LSUN, and Places365 largely contribute to the good mean performance of all OOD detection metrics. We consider the above few datasets are difficult samples in the benchmark since many objects have similar but different semantics. A good result is also achieved on easy datasets of Texture and SVHN.

E. Valuable Comments from Rebuttal

Here posts an answer we highlighted during the rebuttal period to help readers better understand our paper.

[On Motivation of SC-OOD] Classic OOD detection aims to train a 'conservative' model to distinguish samples with either a covariate shift on source distribution $p(x)$ or a semantic shift on label distribution $p(y)$. However, we notice an impractical goal of classic OOD detection: to perfectly distinguish CIFAR cars from ImageNet cars, even though their covariate shift is negligible. The unrealistic goal will unfortunately result in an extremely narrow range of capabilities for deployed models, greatly limiting their use in real applications such as autonomous cars. In an attempt to address this problem, we form a new, realistic, and challenging SC-OOD task that is **juxtaposed** to classic OOD detection. SC-OOD re-defines the 'distribution' as label distribution $p(y)$ only instead of the classic $p(x, y)$. Under the SC-OOD setting, models are required to: 1) well detect images from different label distributions, 2) correctly classify images within the same label distribution with negligible source distribution shifts, which is consistent with a popular research topic called robustness of deep learning.

F. Discussion on Drawbacks

Here we list our current shortcomings. Although the use of UDG mostly helps alleviate the classification decline of the OE method, it can not yet exceed the standard ID classification performance. More exploration is needed for better use of unlabeled data to achieve stronger ID classification while retaining OOD detection capabilities. Also, we will attempt to analyze UDG on larger datasets such as ImageNet with high-resolution images and complex semantics.

Table A2: **Performance details on CIFAR-10 benchmark using ResNet-18.** UDG obtains consistently better results across OOD detection metrics. Accuracy shows the classification accuracy on all the (filtered) ID test samples, which can be improved by UDG on the top of OE method.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				Accuracy ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
MSP	Texture	52.27	90.81	94.07 / 82.32	0.10	1.32	20.84	79.77	95.02
	SVHN	50.25	92.65	87.54 / 95.84	2.47	10.73	48.22	83.96	95.02
	CIFAR-100	61.19	87.40	86.30 / 85.35	0.07	1.72	12.30	69.56	95.02
	Tiny-ImageNet	65.32	87.32	89.41 / 81.17	0.40	2.44	14.16	71.86	92.54
	LSUN	58.62	89.34	89.30 / 86.99	0.88	3.53	19.31	76.46	95.02
	Places365	61.99	87.96	72.61 / 94.64	0.74	2.86	15.63	72.72	93.87
	Mean	**58.27**	**89.25**	**86.54 / 87.72**	**0.78**	**3.77**	**21.74**	**75.72**	**94.42**
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**

Table A3: **Performance details on CIFAR-100 benchmark using ResNet-18.** UDG obtains consistently better results across OOD detection metrics. Accuracy shows the classification accuracy on all the (filtered) ID test samples.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				Accuracy ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
MSP	Texture	84.04	75.85	85.72 / 58.63	0.41	3.67	16.26	45.84	76.65
	SVHN	80.12	80.01	70.84 / 88.52	9.90	17.77	31.00	52.94	76.65
	CIFAR-10	80.64	78.33	80.69 / 74.04	0.00	5.94	21.09	49.10	76.65
	Tiny-ImageNet	83.32	77.85	86.97 / 61.73	2.43	7.55	24.69	48.29	69.56
	LSUN	83.03	77.31	86.31 / 1.45	3.38	6.73	21.49	47.88	76.10
	Places365	77.57	79.99	67.55 / 89.21	1.11	6.02	22.72	51.69	77.56
	Mean	**81.45**	**78.22**	**79.68 / 72.26**	**2.87**	**7.95**	**22.88**	**49.29**	**75.53**
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**

Table A4: **Performance details on CIFAR-10 benchmark using WideResNet-28.** UDG obtains consistently better results across OOD detection metrics. Accuracy shows the classification accuracy on all the (filtered) ID test samples.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				Accuracy ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
MSP	Texture	50.16	89.68	92.45 / 81.81	0.00	0.04	12.16	76.32	96.08
	SVHN	30.54	95.44	92.81 / 97.49	8.75	25.94	72.94	89.16	96.08
	CIFAR-100	51.38	89.15	87.42 / 87.99	0.02	0.77	11.15	75.25	96.08
	Tiny-ImageNet	56.98	88.96	90.14 / 84.19	0.03	0.71	13.85	75.72	93.69
	LSUN	47.05	90.54	88.99 / 89.44	0.20	0.80	11.97	79.25	96.08
	Places365	53.44	89.18	70.65 / 95.54	0.04	0.74	9.22	75.86	95.02
	Mean	**48.26**	**90.49**	**87.08 / 89.41**	**1.51**	**4.83**	**21.88**	**78.59**	**95.51**
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**

Table A5: **Performance details on CIFAR-100 benchmark using WideResNet-28.** UDG obtains consistently better results across OOD detection metrics. Accuracy shows the classification accuracy on all the (filtered) ID test samples, which can be improved by UDG on the top of OE method.

Method	Dataset	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑				Accuracy ↑
					10^{-4}	10^{-3}	10^{-2}	10^{-1}	
MSP	Texture	84.24	76.10	85.25 / 58.36	0.24	2.19	9.78	46.20	80.25
	SVHN	79.63	78.95	65.45 / 88.22	1.42	4.26	17.14	51.39	80.25
	CIFAR-10	77.07	80.81	83.16 / 76.76	0.49	9.19	25.03	53.94	80.25
	Tiny-ImageNet	81.25	79.12	87.75 / 63.33	0.31	5.34	24.75	51.64	72.92
	LSUN	81.32	78.51	86.81 / 62.95	0.51	2.57	20.03	50.74	78.54
	Places365	75.28	80.84	67.81 / 89.76	1.49	4.63	20.12	53.24	80.03
	Mean	**79.80**	**79.05**	**79.37 / 73.23**	**0.74**	**4.70**	**19.48**	**51.19**	**78.71**
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**