8 Supplementary Materials

In the supplementary materials, we provide detailed implementation details and experimental results for both CIFAR and ImageNet benchmarks. Furthermore, we present an analysis to show the effectiveness of our OOD sampling strategy. We then analyze the energy distribution of the ID and OOD samples.

8.1 Additional Implementation Details

For the experiments on CIFAR we use SGD with a momentum of 0.9, and a weight decay of 0.0001. For *GReg* on CIFAR, similar to [36], we decrease the learning rate following a cosine annealing strategy, with a maximum learning rate of 1 and a minimum of 0.001. We use a batch size of 64 and 32 for the CIFAR and ImageNet experiments, respectively.
For both *GReg* and *GReg+* on the ImageNet dataset we perform fine-tuning of a pre-trained DenseNet-121 model (in contrast to the CIFAR benchmarks where we train *GReg+* from scratch) using the ADAM optimizer with an initial learning rate of 10^{-4} and decrease the learning rate to 10^{-5} at epoch 10. We then run *GReg* and *GReg+* to reach epochs 20 and 15, respectively. The other hyperparameters are the same as the CIFAR benchmarks.

To reproduce the results of MSP, ODIN, and Energy, we used the codebase of Energy*. For the case of ReAct and DICE, we utilize the codebase of DICE*, and to reproduce the results of LINe*, and DOS*, we utilize their corresponding codebases. In the spirit of fairness, we run their codes with multiple specifications similar to the original manuscript and report the best results in our tables. Furthermore, experiments requiring training are conducted with 3 different seeds and we report the average values in the tables. All other methods are run with their corresponding default parameters outlined in their manuscript.

8.2 Detailed CIFAR-10/100 Benchmark Results

Table 10 and Table 11 show the complete and detailed performance of various OOD detection approaches on each of the six OOD test datasets. Table 12 compares the sampling methods on the CIFAR-10 dataset. As it can be seen, especially on the WRN and DenseNet architectures, the performance of different methods is saturated.

8.3 Distribution Analysis

To further study the effectiveness of our method, we perform the following analysis. We choose a WRN model pre-trained on CIFAR-10 and fine-tune it with

*https://github.com/wetliu/energy_ood

*https://github.com/deeplearning-wisc/dice

*https://github.com/YongHyun-Ahn/LINe-Out-of-Distribution-Detection-by-Leveraging-Important-Neurons

*https://github.com/lygjwy/DOS

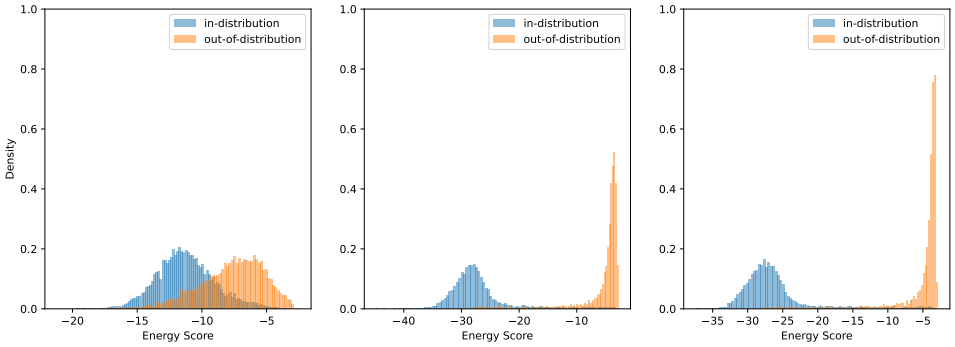

Fig. 4: Distributions of energy scores. The left figure shows the distribution of the pre-trained model, the middle shows the same for Energy loss and the right figure shows the distribution for *GReg*.

Table 7: Ablation study on the number of clusters. FPR95 and AUROC percentages reported on CIFAR benchmarks on DenseNet.

Num. of Clusters	CIFAR-10		CIFAR-100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
No Clustering	**7.93**	**98.12**	56.29	87.35
16	9.88	97.68	58.97	82.88
32	12.1	97.15	46.41	88.433
64	11.27	97.54	**30.55**	**93.38**

the energy loss and *GReg*. Next, we plot the distribution of energy scores for ID (CIFAR-10 test) and OOD (SVHN) datasets and compare the results. As it can be seen from Figure 4, adding gradient regularization to the energy loss enables the network to better distinguish ID samples from the OOD since the distance between the two distributions has increased, i.e., the OOD data are assigned higher energy scores, resulting in further distanced score distributions.

8.4 Sampling

In our setting, we are presented with three main possible clustering spaces: the image space, the feature space, or the logit space. The image space contains the most information but there are two main drawbacks to using it. Firstly, the image space is far too large and high-dimensional, typically in the order of tens of thousands. Secondly, although all the information is in the image, no processing has been done on the image and the features have not been extracted, meaning that in effect, there is also a lot of noise irrelevant to our task. The logit space is not suitable from the opposite perspective, i.e., most of the information has been stripped and only label-relevant information remains. The feature space presents a sweet spot; reasonable dimensionality, relevant features, and smaller levels of noise. Consequently, we choose to cluster the samples in the *feature space*.

Another parameter of interest in the experiments is the number of clusters. In Table 7 we examine the effect of the number of clusters on the performance

of our method on both CIFAR benchmarks using the DenseNet architecture. It can be observed that our sampling significantly improves the results on the CIFAR-100 benchmark, but does not have the same effect on CIFAR-10. Our intuition suggests that one reason behind this observation might be that the CIFAR-10 benchmark is so simple that the DenseNet model can observe all the auxiliary data during the training phase and learn the information. However, in the case of CIFAR-100, as the benchmark is harder, sampling the data helps the model extract more informative samples. Note that the perplexity in comparing CIFAR-100 vs CIFAR-10 is that the datasets contain the same number of samples (60000 total images in both) but CIFAR-100 has 10 times the number of classes of CIFAR-10. This can also be seen in our ImageNet experiments (see Table 3). That is, as the task is harder, sampling using *GReg+* is more effective and improves upon *GReg* by large margins of 12% and 2% on FPR and AUROC, respectively.

8.5 Near-OOD vs Far-OOD Experiments

Apart from the previous experiments, another way to categorize the OOD experiments is by grouping them into Near-OOD (hard-OOD) and Far-OOD (easy-OOD). The intuition behind this grouping is that if the ID data is close to the OOD data, it will be much harder to distinguish between them. Therefore, a powerful OOD method should perform well in both Near-OOD and Far-OOD benchmarks.

To evaluate our performance on the Near-OOD benchmarks, we followed the setup of [75]. In Table 8, we set CIFAR10 as ID and report the metrics on Near-OOD benchmarks (CIFAR100 and Tiny ImageNet) using a ResNet architecture. On average, *GReg+* improves the FPR of DOS by 4% with comparable results in terms of AUROC.

To further compare our method with more auxiliary-based baselines, we use the setup of [75] to perform additional experiments. Table 9, compares our method versus three new benchmarks (MCD [71], UDG [68], MixOE [74]) that use auxiliary data. On average, *GReg+* improves the FPR on CIFAR10 and CIFAR100 by 2.5% and 2.4%, respectively, and achieves comparable AUROC. Note that, unlike DOS, *GReg+* achieves these results with minimal decrease in ID accuracy. Furthermore, *GReg+* outperforms the new benchmarks by large margins.

	Method	CIFAR-100		TIN		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10	ReAct [58]	75.5	85.2	67.61	87.70	71.56	86.47
	Energy score [36]	72.69	85.55	62.41	88.31	67.55	86.93
	Dice [59]	84.29	76.05	75.97	79.53	80.13	77.79
	Energy Loss [36]	48.11	88.24	36.33	91.86	42.22	90.05
	DOS [27]	29.24	**93.38**	16.08	**96.16**	22.66	**94.77**
	GReg+	**23.87**	**93.23**	**13.12**	95.32	**18.49**	**94.28**

Table 8: Near-OOD comparison with CIFAR-10 as ID.

	Method	Near-OOD		Far-OOD		Average		ID Acc
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	
CIFAR-10	MCD [71]	30.17	91.03	32.03	91.00	31.1	91.0	94.9
	UDG [68]	35.34	89.91	20.35	94.06	27.8	91.9	92.3
	MixOE [74]	51.45	88.73	33.84	91.93	42.6	90.3	94.5
	DOS [27]	22.66	94.77	8.01	97.75	15.3	**96.2**	78.5
	GReg+	18.49	94.28	7.47	96.72	**12.9**	95.5	91.1
CIFAR-100	MCD [71]	55.88	77.07	54.39	74.72	55.1	75.8	75.8
	UDG [68]	61.42	78.02	59.00	79.59	60.2	78.8	71.5
	MixOE [74]	55.22	80.95	63.88	76.40	59.5	78.6	75.3
	DOS [27]	56.33	79.63	35.52	87.73	45.9	83.6	47.7
	GReg+	48.26	82.50	38.79	86.4	**43.5**	**84.5**	72.4

Table 9: Additional Experiments on Near-OOD and Far-OOD.

Table 10: Comparison on the CIFAR-10 benchmark. The experimental results are reported over three trials. AUROC and FPR95 are percentages.

Network	Method	OOD Datasets												Average	
		Textures		SVHN		Places		LSUN-c		LSUN-r		iSUN			
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet	MSP [22]	58.78	90.03	73.02	89.42	61.56	87.98	43.88	94.21	53.31	91.38	57.73	89.95	58.04	90.49
	ODIN [34]	50.42	86.89	40.27	91.84	48.65	86.56	7.81	98.21	26.48	94	33.14	92.06	34.46	91.59
	Energy score [36]	53.26	89.12	58.36	91.01	44.51	89.45	12.57	97.64	30.48	94.11	36.77	92.16	39.32	92.24
	ReAct [58]	52.04	89.68	60.2	90.46	44.68	89.26	13.36	97.51	29.95	94.29	36.43	92.63	39.44	92.30
	DICE [59]	54.73	88.34	59.4	89.81	44.09	89.39	22.38	95.72	33.32	93.17	40.52	91.08	42.40	91.25
	LINe [1]	56.97	88.07	66.57	87.4	46.05	88.62	24.9	95.13	34.66	93.12	42.38	91.21	45.25	90.59
	OE [23]	19.46	95.98	13.94	97.22	38.38	90.76	8.15	98.20	21.03	95.87	18.58	96.26	19.92	95.71
	Energy Loss [36]	12.95	97.17	4.22	98.89	27.31	94.05	5.47	98.69	7.87	98.24	9.04	98.16	11.14	97.53
	OpenMix [79]	17.76	96.77	41.12	94.44	27.68	94.4	8.33	98.32	19.06	96.87	19.54	96.79	22.24	96.26
	GReg	5.96	98.36	3	98.83	22.76	94.8	3.03	99	6.5	98.4	6.2	98.33	7.91	97.95
WRN	MSP [22]	59.53	88.45	48.53	91.74	59.86	88.29	31.15	95.6	53.22	91.14	56.87	89.58	51.52	90.8
	ODIN [34]	54.52	80.45	46.6	85.68	48.57	86.04	10.19	97.84	22.86	95.05	28.64	93.87	35.23	89.82
	Energy score [36]	52.52	85.38	36.58	90.73	39.88	89.87	8.2	98.34	28.8	93.87	34.47	92.38	33.40	91.76
	ReAct [58]	53.47	86.58	41.46	89.34	41	90.51	13.21	97.4	34.95	93.30	41.15	91.82	37.54	91.49
	DICE [59]	58.97	84.24	46.09	88.22	42.55	89.69	2.87	99.27	23.6	95.29	30.66	93.78	34.12	91.74
	LINe [1]	57.13	83.81	50.2	83.11	47.35	87.67	2	99.47	27.32	94.32	33.65	92.82	36.27	90.2
	OE [23]	22.56	95.24	27.41	95.13	33.65	92.24	8.88	98.19	17.58	96.22	16.68	96.27	21.12	95.55
	Energy Loss [36]	11.47	97.3	23.93	94.93	22.6	95.37	5.07	98.83	8.07	98.17	7.53	98.27	13.11	97.14
	OpenMix [79]	21.17	95.85	26.63	95.65	27.17	94.13	14.32	97.35	21.8	96.42	20.45	96.61	21.92	96
	GReg	7.83	98.16	6.23	98.4	19.9	95.66	3.8	98.93	5.26	98.7	4.66	98.76	7.95	98.10
DenseNet	MSP [22]	66.3	87.09	44.64	93.86	63.16	88.35	43.34	94.17	48.9	93.4	49.05	93.35	52.56	91.70
	ODIN [34]	55.62	85.03	29.03	94.86	42.44	91.11	11.94	97.67	4.86	98.96	5.39	98.9	24.88	94.42
	Energy score [36]	60.03	85.17	35.2	94.76	40.9	91.58	9.5	98.17	13.78	97.51	14.57	97.4	28.99	94.09
	ReAct [58]	50.47	90.53	29.81	95.56	40.35	92.05	10.04	98.11	11.44	97.79	12.92	97.62	25.83	95.27
	DICE [59]	56.26	86.42	40.47	93.99	39.6	91.82	3.79	99.15	8.71	98.19	9.42	98.11	26.37	94.61
	LINe [1]	23.35	95.11	12.22	97.56	43.72	91.13	0.61	99.83	4.09	99.1	5.06	99.02	14.84	96.95
	OE [23]	20.93	96.04	12.19	97.68	39.18	91.69	7.15	98.57	25.45	95.42	25.72	95.41	21.77	95.8
	Energy Loss [36]	13.43	97.04	20.08	95.52	20.09	95.41	3.53	99.19	4.55	98.84	5.87	98.63	11.26	97.44
	OpenMix [79]	24.8	95.3	40.59	92.58	29.13	93.88	12.39	97.64	15.4	97.23	14.9	97.31	22.87	95.66
	GReg	8.3	97.9	3.7	98.83	20.13	95.66	3.1	99.23	6.36	98.53	6.03	98.56	7.93	98.12

Table 11: Comparison on the CIFAR-100 benchmark. The experimental results are reported over three trials. AUROC and FPR95 are percentages.

Network	Method	OOD Datasets												Average	
		Textures		SVHN		Places		LSUN-c		LSUN-r		iSUN			
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet	MSP [22]	82.18	75.44	70.19	82.53	82.09	75.06	64.37	85.43	74.74	81.24	74.39	80.94	74.66	80.1
	ODIN [34]	76.44	77.59	65.24	84.61	83.37	74.62	41.63	93.07	59.85	88	57.48	88.09	64	84.33
	Energy score [36]	81.82	77.71	57.08	89.09	84.12	75.4	50.24	91.71	69.73	85.49	67.77	85.5	68.46	84.15
	ReAct [58]	67.69	85.5	48.94	91.8	82.29	77.2	31.54	94.61	63.9	88.53	61.96	87.95	59.38	87.59
	DICE [59]	88.92	71.5	69.78	85.12	87.81	72.13	59.78	89.08	77.69	82.87	79.81	81.42	77.29	80.35
	LINe [1]	78.71	79.96	66.2	86.98	87.71	72.88	43.98	91.09	78.27	83.48	77.69	83.19	72.09	82.93
	OE [23]	73.91	73.14	86.45	70.18	69.28	77.75	44.44	86.51	94.55	38.80	95.30	36.98	77.32	63.89
	Energy Loss [36]	57.3	83.9	75.06	84.33	82.2	75.8	35.03	93.93	62.06	83.4	65	82.93	62.77	84.05
	POEM [43]	67.92	77.69	78.72	78.6	89.07	68.76	41.97	92.24	46.27	88.31	42.72	89.01	61.11	82.43
	OpenMix [79]	59.87	87.54	72.91	86.76	75.73	80	48.75	91.31	48.02	92.25	52.56	90.71	59.64	88.09
	DOS [27]	52.29	88.1	47.57	91.44	77.99	81.42	54.75	89	45.5	90.76	49.62	89.08	54.62	88.30
	GReg	58.73	83.73	42	91.7	78.4	76.7	28.56	94.6	74.56	75.2	75.33	75.63	59.6	82.92
	GReg+	48.74	88.2	42.2	92.45	69.61	83.35	43.86	90.75	46.04	89.94	54.22	87.8	50.78	88.75
WRN	MSP [22]	83.05	73.75	83.88	71.97	82.03	73.91	66.59	83.71	83.11	74.12	83.66	74.77	80.39	75.37
	ODIN [34]	76.57	75.13	90.37	67.08	81.12	74.21	36.44	93.32	62	84.55	61.23	84.89	67.95	79.86
	Energy score [36]	79.78	76.47	85.57	74.43	80.07	75.69	36.09	93.4	80.6	77.72	82.82	77.86	74.15	79.26
	ReAct [58]	67.12	82.97	74.45	88.28	81.21	74.24	36.63	91.67	81.23	72	83.31	72.07	70.65	80.20
	DICE [59]	85.32	72.84	87.21	71.39	81.02	74.95	16.73	96.83	79.86	81.64	83.82	80.08	72.32	79.62
	LINe [1]	67.34	81.98	81.73	83.6	84.03	71.03	16.93	96.61	77.98	76.26	77	76.37	67.50	80.97
	OE [23]	73.24	72.47	71.6	81.3	69.2	77.97	40.46	87.26	95.04	38.36	95.77	36.84	74.22	65.7
	Energy Loss [36]	68.96	80.66	87.7	73.6	83.26	74.36	55.53	88.53	47.7	81.96	50.6	82	65.62	80.18
	POEM [43]	79.23	70.17	90.37	67.89	85.89	70.51	26.09	94.35	45.3	88.59	48.59	86.92	62.58	79.74
	OpenMix [79]	64.91	83.61	87.25	70.34	73.17	78.14	59.18	87.77	75.19	78.46	79.02	77.05	73.12	79.23
	DOS [27]	41.89	91.58	15.07	97.33	59.18	88.24	32.12	94.38	61.31	86.78	62	86.22	45.26	90.76
	GReg	54.5	85.53	57.33	89.76	74.36	79.53	43.86	91.73	58.33	86.8	61.16	86.03	58.26	86.56
	GReg+	51.6	90	25.73	95.96	65.16	87.56	41.68	93.42	51.61	89.96	52.93	89.24	48.12	91.02
DenseNet	MSP [22]	86.12	70.66	85.55	72.33	83.25	74.01	77.45	78	92.18	57.09	90.71	60.59	85.87	68.78
	ODIN [34]	80.14	73.74	80.63	83.75	76.78	78.75	43.87	91.25	69.6	80.44	68.8	81.4	69.97	81.55
	Energy score [36]	84.81	70.29	88.6	80.99	78.12	77.48	51.08	88.73	84.56	69.31	86.3	69.96	78.91	76.12
	ReAct [58]	78.22	77.9	88.02	78.76	81.98	73.53	53.51	88.13	76.23	81.32	79.7	80.97	76.27	80.10
	DICE [59]	84.34	71.02	87.51	81.86	78.19	78.15	14.75	97.44	75.36	77.77	78.7	76.8	69.80	80.50
	LINe [1]	39.24	87.91	31.6	91.7	88.48	63.82	5.75	98.85	23.33	94.95	22.6	95.13	35.16	88.72
	OE [23]	74.55	77.07	61.17	87.34	66.75	81.56	24.98	95.1	65.02	83.12	70.65	81.11	60.52	84.21
	Energy Loss [36]	70.03	81.3	70.78	88.42	62.45	85.15	26.09	95.38	77.59	75.23	79.32	77.69	64.37	83.86
	POEM [43]	75.48	73.04	83.57	71.49	83.6	73.99	33.96	93.43	39.74	83.58	39.61	83.33	59.33	79.81
	OpenMix [79]	63.66	84.05	72.27	85.77	73.17	80.19	48.79	90.8	71.76	83.77	74.1	82.23	67.29	84.47
	DOS [27]	38.3	91.72	11.57	97.88	57.06	88.91	25.32	95.58	38.43	92.75	38.85	92.49	34.92	93.22
	GReg	64.63	81.74	45.79	92.53	78.1	78.1	31.27	94.53	55.98	89.43	61.96	87.8	56.29	87.35
	GReg+	41.83	90.19	14.44	97.06	53.56	89.11	19.49	95.8	26.01	94.36	27.95	93.77	30.55	93.38

Table 12: Comparison of methods with sampling on the CIFAR-10 benchmark. The experimental results are reported over three trials. AUROC and FPR95 are percentages.

Network	Method	OOD Datasets												Average	
		Textures		SVHN		Places		LSUN-c		LSUN-r		iSUN			
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet	POEM [43]	33.24	93.51	58.95	89.58	61.82	86.26	62.78	90.13	11.22	97.96	10.53	98.03	39.76	92.58
	DOS [27]	7.76	98.74	4.82	99.19	21.3	95.50	9.68	98.39	10.86	97.93	12.31	97.68	11.12	97.91
	GReg	5.96	98.36	3	98.83	22.76	94.8	3.03	99	6.5	98.4	6.2	98.33	**7.91**	**97.95**
	GReg+	7.70	98.47	5.65	98.79	24.25	94.80	9.54	98.18	14.56	97.30	15.04	97.11	12.79	97.44
WRN	POEM [43]	31.08	93.84	64.09	87.23	57.86	87.62	29.51	95.01	25.97	95.08	23.17	95.48	38.61	92.38
	DOS [27]	4.33	99	1.78	99.32	10.89	97.16	3.73	98.98	9.18	98.2	9.86	98.17	6.63	**98.47**
	GReg	7.83	98.16	6.23	98.4	19.9	95.66	3.8	98.93	5.26	98.7	4.66	98.76	7.95	98.10
	GReg+	4.19	98.52	2.92	98.59	11.38	96.71	3.88	98.63	5.63	98.47	6.57	98.27	**5.76**	98.2
DenseNet	POEM [43]	36.79	91.31	31.69	94.34	53.69	88.34	43.57	92.84	5.06	99.03	4.11	99.23	29.15	94.18
	DOS [27]	2.61	99.39	0.67	99.65	7.16	97.90	0.88	99.56	1.61	99.21	1.53	99.27	**2.41**	**99.16**
	GReg	8.3	97.9	3.7	98.83	20.13	95.66	3.1	99.23	6.36	98.53	6.03	98.56	7.93	98.12
	GReg+	6.21	98.29	2.19	98.73	21.5	95.1	6.87	98.27	15.53	97.45	15.36	97.43	11.27	97.54