Appendix

A Smaller Perturbation Budget

In Table 1, we reported the AUROC for the OOD detection for the perturbed in/out distributions by the PGD attack with $\epsilon = \frac{8}{255}$. Here, a similar evaluation is conducted with half the perturbation budget. The results are displayed in Table 6. Similar to Table 1, HAT is the best method among the previous defenses, and ATD outperforms it with a significant margin. ATD achieves satisfactory robustness in all of the attack settings in both datasets, along with a decent clean detection AUROC score.

It should be noted that attacking both in and out sets is included in the experiments to give a sense of how the distribution of OOD scores changes under adversarial attacks, but evaluating the AUROC when perturbing either in or out set is more reasonable in practice.

Table 6: OOD detection AUROC under PGD attack with $\epsilon = \frac{4}{255}$ for various methods trained with CIFAR-10 or CIFAR-100 as the closed set. A clean evaluation is one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked. The best and second-best results are distinguished with bold and underlined text for each column.

Method	In-Distribution Dataset							
	CIFAR-10				CIFAR-100			
	Clean	In	Out	In and Out	Clean	In	Out	In and Out
OpenGAN-fea	0.971	0.521	0.444	0.318	**0.958**	0.283	0.382	0.123
OpenGAN-pixel	0.818	0.001	0.052	0.000	0.767	0.001	0.026	0.000
ViT (MSP)	0.975	0.578	0.280	0.013	0.879	0.362	0.157	0.005
ViT (MD)	**0.995**	0.260	0.571	0.004	<u>0.951</u>	0.094	0.327	0.001
ViT (RMD)	0.951	0.526	0.493	0.043	0.915	0.366	0.371	0.041
ViT (OpenMax)	<u>0.984</u>	0.339	0.355	0.007	0.907	0.164	0.205	0.003
AT (MSP)	0.735	0.573	0.564	0.394	0.603	0.416	0.375	0.226
AT (MD)	0.771	0.603	0.650	0.473	0.649	0.454	0.495	0.310
AT (RMD)	0.836	0.640	0.690	0.453	0.700	0.512	0.511	0.329
AT (OpenMax)	0.805	0.647	0.669	0.489	0.650	0.469	0.491	0.326
HAT(MSP)	0.770	0.665	0.658	0.541	0.612	0.484	0.457	0.339
HAT(MD)	0.789	0.691	0.688	0.579	0.810	<u>0.712</u>	<u>0.711</u>	<u>0.596</u>
HAT (RMD)	0.878	<u>0.740</u>	<u>0.764</u>	0.573	0.730	0.579	0.562	0.402
HAT (OpenMax)	0.821	0.729	0.741	<u>0.633</u>	0.703	0.587	0.586	0.464
OSAD (MSP)	0.698	0.521	0.516	0.347	0.557	0.346	0.295	0.157
OSAD (MD)	0.626	0.500	0.521	0.402	0.615	0.491	0.510	0.389
OSAD (RMD)	0.776	0.576	0.619	0.384	0.680	0.500	0.495	0.321
OSAD (OpenMax)	0.827	0.696	0.699	0.535	0.647	0.476	0.478	0.318
AOE (MSP)	0.780	0.658	0.654	0.514	0.566	0.435	0.430	0.313
AOE (MD)	0.709	0.515	0.587	0.407	0.743	0.535	0.636	0.437
AOE (RMD)	0.780	0.562	0.595	0.337	0.682	0.464	0.468	0.270
AOE (OpenMax)	0.797	0.675	0.702	0.559	0.591	0.427	0.468	0.315
ALOE (MSP)	0.843	0.667	0.612	0.373	0.701	0.489	0.440	0.258
ALOE (MD)	0.827	0.602	0.646	0.406	0.793	0.623	0.656	0.466
ALOE (RMD)	0.815	0.512	0.570	0.213	0.632	0.398	0.402	0.207
ALOE (OpenMax)	0.869	0.658	0.679	0.422	0.731	0.483	0.509	0.282
ATD (Ours)	0.943	**0.915**	**0.919**	**0.883**	0.877	**0.818**	**0.817**	**0.742**

Table 7: Comparing OOD detection methods robustness. For each of MSP, MD, RMD, and OpenMax, the AUROC is averaged over six different baselines in Table 1 that use these detection methods. A clean evaluation is the one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked.

Method	In-Distribution Dataset							
	CIFAR-10				CIFAR-100			
	Clean	In	Out	In and Out	Clean	In	Out	In and Out
MSP	0.800	0.515	0.439	0.204	0.653	0.340	0.258	0.100
MD	0.786	0.374	0.500	0.196	**0.760**	**0.368**	**0.456**	**0.201**
RMD	0.839	0.427	0.469	0.109	0.723	0.362	0.348	0.117
OpenMax	**0.850**	**0.514**	**0.532**	**0.239**	0.705	0.312	0.342	0.131

Figure 5: Images from the open and closed sets before and after the attack with $\epsilon = \frac{8}{255}$. HAT+MSP is used as the base model in attacking the images. The images are perceptually similar before and after the attack. Also, the OOD detection score for each image is displayed above it, which demonstrates the effectiveness of the attack.

B Detection Method

To investigate the effect of the detection method, we averaged the results in Table 1 over all the baselines (ViT, AT, HAT, OSAD, AOE, and ALOE) for MSP, MD, RMD, and OpenMax. According to the results in Table 7, OpenMax and MD are the best choices when considering CIFAR-10 and CIFAR-100 as the closed set, respectively.

C Visualization of Adversarial Images

A number of the original and adversarial images for the open and closed sets are shown in Fig. 5. Clearly, images have not changed perceptually under the attack with $\epsilon = \frac{8}{255}$, while the detection score has changed significantly.

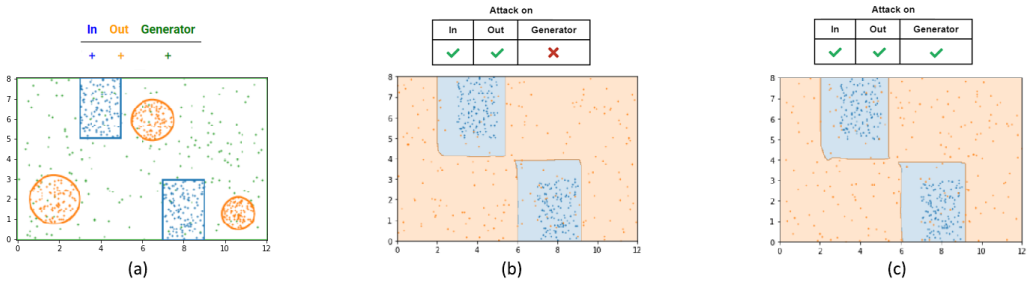

Figure 6: Effect of attacking generated data along with the in and out sets.

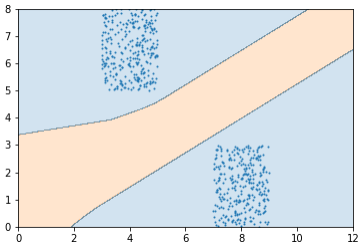

Figure 7: OOD detection with an adversarially trained model using MSP as the detection method. Each of the rectangular distributions are regarded as a separate class, and adversarial training is conducted on them. Next, OOD scores are extracted with the MSP method all over the space, and thresholded to achieve the in/out decision boundaries. The blue and orange backgrounds are regarded as the in- and out-distributions, respectively.

D Toy Example Exploration

In section 3.2, a toy example was investigated with different attack settings. Here, we use this toy example to investigate two other configurations. First, we use a similar setting to check the effect of attacking the generated data along with the closed and open sets. The results in Fig. 6 demonstrate that attacking the generated data during training does not change the decision boundaries with the setting in this example. As mentioned earlier, we believe that attacking the generator should not be done in practice due to the inevitable instability in training of the GAN.

In the second configuration, we use this toy example to simulate the pure adversarial training with MSP detection method in OOD detection. To this end, we only consider the in-distribution data in training. Each of the blue rectangles in Fig. 7 are regarded as a separate class, and adversarial training is performed on them. Next, MSP is used as the score function all over the considered space. Finally, the scores are separated with a threshold to detect OOD samples. The in/out classification decision boundaries are displayed in Fig. 7, which demonstrates that pure AT is not as effective as ATD for in/out robust classification due to the lack of OOD samples in the training.

E Number of Attack Steps

To check the effect of increasing the attack steps, we plotted the AUROC score with various attack steps for HAT, OSAD, and ATD. According to Fig. 8, 10 steps are enough for evaluation of HAT and OSAD, and 50 steps are enough for ATD. Note that we have used attack with 100 steps in the evaluations of ATD to ensure its robustness.

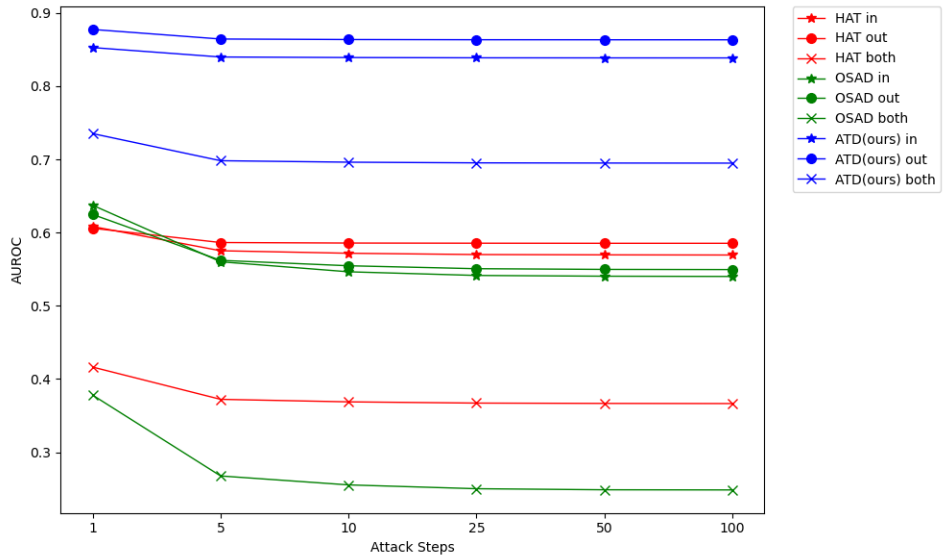

Figure 8: Effect of the number of PGD attack steps on the detection AUROC in HAT, OSAD, and ATD methods trained with CIFAR-10 as the closed set. Three attacking scenarios are considered: attacking only the in data, attacking only the out data, and attacking both in and out data. Attacks are conducted with 1, 5, 10, 25, 50, and 100 steps. The AUROC reaches a stable value after 50 steps for all the methods and scenarios.

Table 8: Comparison of AUROC score for OOD detection with ATD method under PGD-100 attack and AutoAttack with $\epsilon = \frac{8}{255}$. CIFAR-10 and CIFAR-100 datasets are used as the in-distribution dataset. A clean evaluation is one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked.

Attack	In-Distribution Dataset							
	CIFAR-10				CIFAR-100			
	Clean	In	Out	In and Out	Clean	In	Out	In and Out
PGD-100	0.943	0.837	0.862	0.693	0.877	0.734	0.739	0.553
Auto Attack	0.943	0.834	0.861	0.688	0.877	0.729	0.737	0.545

F Evaluation With AutoAttack

Despite the high power of the PGD attack in assessing the robustness, new attacks have been proposed recently that aim to better evaluate the robustness with stronger attacks. Among them, AutoAttack [64] has become a more reliable substitute to PGD and have proven to be much more difficult to be fooled. AutoAttack propose to use an ensemble of four diverse attacks to reliably evaluate the robustness. These attacks are run sequentially on an input batch, and the one that causes the most loss is used for perturbing the input data. To ensure the reliability of our results, we have evaluated our method against such a strong attack with $\epsilon = \frac{8}{255}$ in addition to PGD-100 in the Table 8. The results indicate that the model is sufficiently robust against such a strong attack, and the results are still promising.

G TinyImageNet Dataset

Adversarial robustness on larger and sophisticated datasets is still a challenging issue, even in closed set classification. For instance, the adversarial accuracy on the closed set in the TinyImageNet dataset ($\epsilon = \frac{8}{255}$) is less than 20% [46]. On the other hand, closed set accuracy significantly affects OOD

Table 9: OOD detection AUROC under PGD attack with $\epsilon = \frac{4}{255}$ for ViT, AT, HAT, ALOE, and ATD trained with TinyImageNet as the closed set. A clean evaluation is one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked.

Method	Clean	In	Out	In and Out
ViT (MSP)	0.924	0.229	0.040	0.000
ViT (MD)	**0.954**	0.198	0.208	0.000
ViT (RMD)	0.934	0.204	0.250	0.001
ViT (OpenMax)	0.915	0.216	0.031	0.000
AT (MSP)	0.557	0.313	0.309	0.127
AT (MD)	0.397	0.160	0.208	0.055
AT (RMD)	0.522	0.254	0.263	0.086
AT (OpenMax)	0.536	0.272	0.305	0.112
HAT (MSP)	0.563	0.377	0.376	0.200
HAT (MD)	0.676	0.498	0.495	0.309
HAT (RMD)	0.710	0.479	0.419	0.205
HAT (OpenMax)	0.574	0.381	0.385	0.205
ALOE (MSP)	0.575	0.233	0.080	0.005
ALOE (MD)	0.689	0.083	0.350	0.002
ALOE (RMD)	0.500	0.136	0.087	0.008
ALOE (OpenMax)	0.569	0.114	0.120	0.002
ATD (Ours)	0.883	**0.786**	**0.781**	**0.635**

detection [11]. As a result, adversarially robust OOD detection on large-scale datasets would be much harder since the robust closed set classification itself is still an open challenging issue.

Still, we run an experiment on TinyImageNet as the in-distribution dataset that compares our method with ViT, ALOE, HAT, and AT methods. AT, HAT, and ALOE are evaluated against PGD-10 with $\epsilon = \frac{4}{255}$, while PGD-100 with $\epsilon = \frac{4}{255}$ is used for ViT and ATD. The training hyper-parameters are similar to the setup in section 4.1. Also, TinyImageNet is excluded from the OOD datasets. The results are listed in Table 9 that demonstrate our method's robustness.

H Certified Detection

Recently, several works have tried to tackle the issue of unreliable evaluations of robustness in neural networks by providing provable guarantees. This is not limited to OOD detection, and it was first noticed in image classification methods. For instance, a provable bound on the confidence score of neural network was provided for image classification methods using randomized smoothing method [65], or a lower and upper bound was provided by IBP [66] for the output of each layer in the neural network given that input x is varied in the ℓ_∞ ball of radius ϵ. Similar efforts have been made in the OOD detection field. For instance, GOOD [67] and ProoD [68] use a similar method to IBP to provide certified bounds.

Despite the advances in the certified robustness research field, the certified defenses still cannot compete with the adversarially trained models against commonly perturbation budget $\epsilon = \frac{8}{255}$ since their goal is to provide provable bounds in specific conditions based on the input itself and the adversary perturbation budget, and the model is not optimized only on the adversarial examples.

I OOD Detection Details

In the experiments, CIFAR-10 and CIFAR-100 are considered the in-distribution datasets, while OOD datasets include MNIST, TinyImageNet, Places365, LSUN, iSUN, Birds, Flowers, and COIL-100. Next, the results are averaged over all the OOD datasets and reported in Table 1.

In this section, the details of the experiments are provided for each OOD dataset in the Tables 10 and 11 for CIFAR-10 and CIFAR-100, respectively. For the baseline methods which are evaluated

with different detection methods such as MSP, MD, RMD, and Openmax, only the one with the best average across all the "Clean", "In", "Out", "In and Out" cases is chosen. Based on these tables, our method is more robust than other methods, even in a single OOD dataset in addition to the average case. Hence, we can conclude that our method is more robust than other baselines across a wide range of OOD datasets.

Note that some of the OOD datasets are somewhat similar to the in-distribution data (e.g. Birds and the bird class from CIFAR-10). These datasets can be regarded as the near-OOD data that the model is expected to detect them as well as the other OODs. Therefore, averaging across all these datasets helps to evaluate the models even on the near-OOD data [69]. In addition, in Tables 10 and 11, CIFAR-10 and CIFAR-100 are also considered as the OOD dataset for the other one, which can be regarded as another case of the near-OOD study [70].

J Computational Cost

Using a single RTX 2060 Super GPU and the setup mentioned in the Section 4.1, our model training and evaluation last for 5 and 2.75 hours, respectively. The time for pretraining the feature extractor model is not included in the mentioned times, which will last about 9 hours in total. Therefore, training ATD would be feasible in a reasonable time, even for large datasets. The evaluation time of the other baseline methods is also similar when using the same setting and MSP detector. This is because they all need to obtain features from the image through a convolutional backbone, which takes most of the evaluation time. Furthermore, when using distance-based detectors for the baseline methods, the evaluation time will increase by five times since they require fitting a class conditional distribution to the pre-logit features.

Table 10: OOD detection AUROC for each of OOD datasets under PGD attack with $\epsilon = \frac{8}{255}$ for various methods trained with CIFAR-10 as the closed set. A clean evaluation is one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked.

Method	Attacked Distribution	Out-Distribution Dataset								
		MNIST	TiImgNet	Places	LSUN	iSUN	Birds	Flower	COIL	CIFAR100
OpenGAN-fea	Clean	0.994	0.953	0.950	0.965	0.963	0.983	0.983	0.981	0.950
	In	0.631	0.363	0.392	0.434	0.425	0.520	0.489	0.527	0.316
	Out	0.486	0.433	0.363	0.420	0.419	0.408	0.387	0.481	0.296
	In and Out	0.303	0.152	0.173	0.240	0.230	0.345	0.301	0.385	0.101
ViT (RMD)	Clean	0.987	0.952	0.983	0.984	0.986	0.760	0.996	0.959	0.973
	In	0.393	0.457	0.491	0.551	0.544	0.099	0.452	0.429	0.443
	Out	0.550	0.408	0.481	0.401	0.396	0.235	0.480	0.621	0.364
	In and Out	0.035	0.023	0.031	0.020	0.021	0.006	0.026	0.039	0.017
AT (OpenMax)	Clean	0.804	0.810	0.825	0.850	0.839	0.751	0.855	0.703	0.796
	In	0.488	0.460	0.482	0.507	0.495	0.403	0.512	0.397	0.444
	Out	0.729	0.455	0.493	0.509	0.507	0.419	0.530	0.426	0.439
	In and Out	0.396	0.165	0.189	0.196	0.195	0.147	0.209	0.166	0.160
HAT (OpenMax)	Clean	0.750	0.831	0.847	0.880	0.857	0.741	0.888	0.776	0.805
	In	0.538	0.616	0.641	0.677	0.650	0.516	0.690	0.573	0.579
	Out	0.669	0.647	0.678	0.707	0.679	0.537	0.687	0.581	0.606
	In and Out	0.452	0.406	0.438	0.458	0.435	0.320	0.441	0.367	0.368
OSAD (OpenMax)	Clean	0.862	0.819	0.833	0.864	0.840	0.765	0.886	0.750	0.799
	In	0.665	0.498	0.538	0.566	0.532	0.419	0.653	0.475	0.474
	Out	0.794	0.506	0.528	0.533	0.525	0.504	0.574	0.469	0.484
	In and Out	0.553	0.193	0.212	0.207	0.203	0.191	0.266	0.187	0.181
AOE (OpenMax)	Clean	0.584	0.820	0.877	0.922	0.902	0.798	0.723	0.749	0.782
	In	0.287	0.539	0.620	0.680	0.648	0.528	0.419	0.500	0.492
	Out	0.496	0.571	0.650	0.699	0.670	0.575	0.456	0.571	0.530
	In and Out	0.225	0.283	0.345	0.380	0.357	0.287	0.198	0.310	0.252
ALOE (MSP)	Clean	0.746	0.821	0.851	0.987	0.983	0.799	0.790	0.768	0.788
	In	0.463	0.609	0.661	0.936	0.922	0.591	0.585	0.541	0.458
	Out	0.521	0.470	0.478	0.757	0.743	0.463	0.437	0.434	0.476
	In and Out	0.227	0.216	0.228	0.516	0.504	0.218	0.196	0.193	0.170
ATD (Ours)	Clean	0.988	0.880	0.925	0.960	0.948	0.936	0.997	0.908	0.820
	In	0.938	0.685	0.795	0.861	0.840	0.839	0.983	0.757	0.580
	Out	0.977	0.726	0.813	0.879	0.857	0.853	0.985	0.808	0.641
	In and Out	0.902	0.470	0.607	0.690	0.668	0.690	0.937	0.581	0.380

Table 11: OOD detection AUROC for each of OOD datasets under PGD attack with $\epsilon = \frac{8}{255}$ for various methods trained with CIFAR-100 as the closed set. A clean evaluation is one where no attack is made on the data, whereas an in/out evaluation means that the corresponding data is attacked.

Method	Attacked Distribution	Out-Distribution Dataset								
		MNIST	TiImgNet	Places	LSUN	iSUN	Birds	Flower	COIL	CIFAR10
OpenGAN-fea	Clean	0.990	0.883	0.945	0.971	0.964	0.966	0.968	0.977	0.929
	In	0.267	0.134	0.169	0.217	0.205	0.195	0.178	0.219	0.299
	Out	0.493	0.149	0.223	0.331	0.347	0.268	0.347	0.438	0.276
	In and Out	0.146	0.039	0.049	0.073	0.075	0.074	0.093	0.157	0.091
ViT (RMD)	Clean	0.838	0.901	0.923	0.916	0.914	0.978	0.966	0.881	0.948
	In	0.271	0.359	0.325	0.340	0.336	0.516	0.495	0.280	0.386
	Out	0.192	0.335	0.419	0.199	0.252	0.687	0.454	0.351	0.485
	In and Out	0.017	0.031	0.037	0.006	0.009	0.105	0.055	0.035	0.058
AT (RMD)	Clean	0.411	0.723	0.731	0.760	0.725	0.731	0.776	0.744	0.675
	In	0.178	0.369	0.386	0.404	0.375	0.373	0.449	0.396	0.320
	Out	0.353	0.329	0.347	0.354	0.326	0.360	0.411	0.425	0.295
	In and Out	0.142	0.120	0.127	0.129	0.119	0.134	0.157	0.163	0.107
HAT (MD)	Clean	0.992	0.662	0.780	0.826	0.787	0.829	0.879	0.723	0.540
	In	0.962	0.379	0.521	0.583	0.531	0.598	0.676	0.444	0.268
	Out	0.988	0.389	0.524	0.580	0.523	0.585	0.708	0.525	0.291
	In and Out	0.943	0.161	0.253	0.292	0.254	0.318	0.429	0.254	0.108
OSAD (MD)	Clean	0.959	0.483	0.557	0.556	0.548	0.545	0.696	0.575	0.503
	In	0.865	0.239	0.293	0.279	0.280	0.270	0.429	0.286	0.261
	Out	0.942	0.266	0.320	0.306	0.299	0.313	0.494	0.384	0.277
	In and Out	0.820	0.099	0.121	0.104	0.106	0.110	0.229	0.140	0.103
AOE (MD)	Clean	0.982	0.540	0.687	0.740	0.724	0.747	0.837	0.690	0.443
	In	0.715	0.211	0.332	0.386	0.385	0.418	0.493	0.312	0.149
	Out	0.969	0.299	0.428	0.482	0.460	0.534	0.669	0.472	0.236
	In and Out	0.684	0.096	0.151	0.192	0.187	0.255	0.322	0.154	0.061
ALOE (MD)	Clean	0.966	0.581	0.750	0.831	0.801	0.784	0.851	0.779	0.436
	In	0.806	0.235	0.424	0.596	0.557	0.503	0.579	0.430	0.131
	Out	0.947	0.265	0.439	0.506	0.488	0.513	0.647	0.535	0.168
	In and Out	0.731	0.063	0.141	0.207	0.221	0.237	0.318	0.192	0.030
ATD (Ours)	Clean	0.973	0.737	0.833	0.892	0.865	0.934	0.972	0.806	0.575
	In	0.903	0.497	0.659	0.735	0.704	0.835	0.921	0.619	0.320
	Out	0.959	0.492	0.636	0.724	0.688	0.834	0.909	0.667	0.322
	In and Out	0.863	0.260	0.417	0.494	0.473	0.662	0.801	0.453	0.138