A Non-robust Model Training

For training, *CIFAR-10/100* data was zero-padded by 4 px along each dimension, and then transformed using 32×32 px random crops, and random horizontal flips. Channel-wise normalization was replicated as reported by the original dataset authors. Training hyper parameters have been set to an initial learning rate of 1e-2, a weight decay of 1e-2, a batch-size of 256 and a nesterov momentum of 0.9. We scheduled the SGD optimizer to decrease the learning rate every 30 epochs by a factor of $\gamma = 0.1$ and trained for a total of 125 epochs. The loss is determined using Categorical Cross Entropy and we used the model obtained at the epoch with the highest validation accuracy. Training was executed on a *A+ Server* SYS-2123GQ-NART-2U machine with four *NVIDIA* A100-SXM4-40GB GPUs for approximately 17 GPU hours. Training *ImageNet1k* architectures with our hyperparameters resulted in a rather poor performance and we therefore rely on the baseline model without AT provided by *timm* [77].

B Additional Evaluation CIFAR10/100

In this section we provide an overview over ECE on CIFAR10 and CIFAR100 of all robust models and their non-robust counterparts.

B.1 Confidence Distribution

The model confidence distributions are shown in Figure 9 and Figure 10. Each row contains the robust and non-robust counterpart and their confidence distributions on the clean samples and the perturbated samples by PGD and Squares.

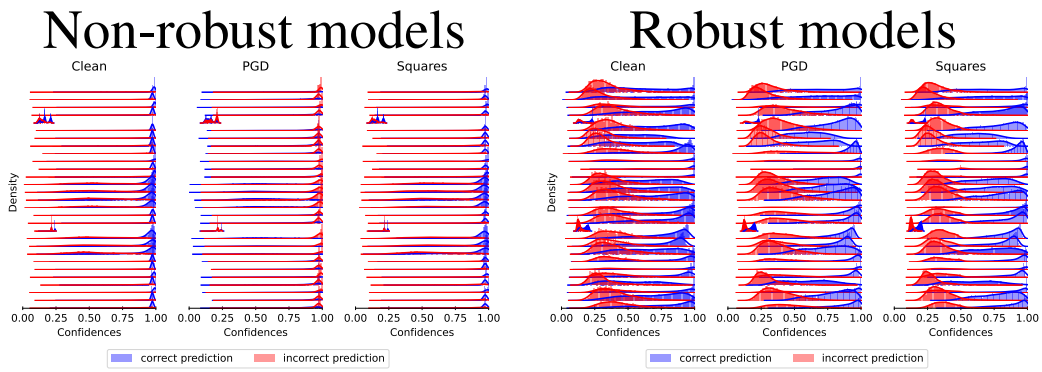

Figure 9: Density plots for robust and non-robust models on CIFAR10 over the models confidence on its correct and incorrect predictions. Each row contains the same model adversarially and standard trained. The non-robust models show high confidence in all of their predictions, however, those might be wrong. Especially in the case of PGD samples, the models are highly confident in their false predictions. In contrast, the robust models are better calibrated. The robust models are confident in their correct predictions and less confident in their false predictions.

B.2 Overconfidence and ECE

Similar, the confidence distributions for the robust and non-robust counterparts on CIFAR100 are depicted in Figure 10.

Samples \ Robustness	Clean	PGD	Squares
non-robust models	0.3077 ± 0.1257	0.2159 ± 0.0738	0.2780 ± 0.1348
robust models	0.2962 ±0.1722	0.2307 ± 0.1494	0.2076 ± 0.1247

Table 2: Mean ECE (lower is better) and standard deviation over all non-robust model versus all their robust counterparts trained on CIFAR100. Robust model exhibit a significantly lower ECE on all samples.

B.3 Precision Recall

For completeness, we included the Precision Recall curves on CIFAR10

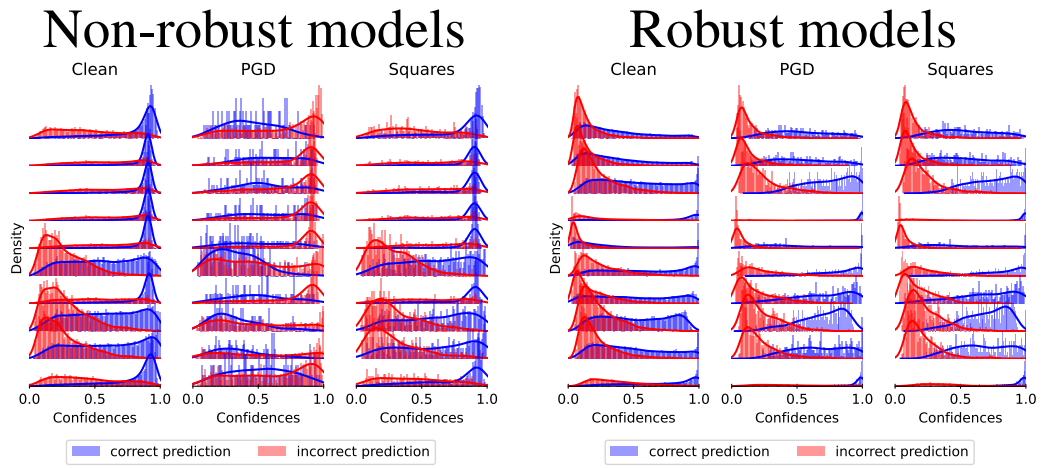

Figure 10: Density plots for robust and non-robust models on CIFAR100 over the models confidence on its correct and incorrect predictions. Each row contains the same model adversarially and standard trained. The non-robust models show high confidence in all of their predictions, however, those might be wrong. Especially in the case of PGD samples, the models are highly confident in their false predictions. In contrast, the robust models are better calibrated. The robust models are confident in their correct predictions and less confident in their false predictions.

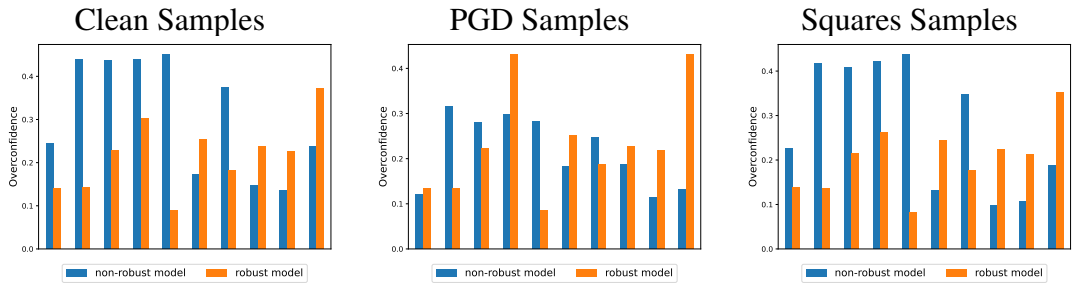

Figure 11: Overconfidence (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR100.

and CIFAR100 as mean over all robust and non-robust models with marked standard deviation.

C FLC Pooling

We evaluate different robust PRN-18 networks trained with flc pooling [27] and FGSM AT in terms of their confidence distribution. For training, we used the training script provided by [78]. We trained with ten different seeds and run for 300 epochs, choosing the batchsize to be 128, a momentum of 0.9, weight decay of 0.0005, a cycling learning rate with minimum value of 0 and maximum value of 0.2, for the adversarial samples we used FGSM with an ϵ of $8/255$ and α of $10/255$. Figure 18 shows the confidence distribution over all ten models and the standard deviation between those models. We can observe that the models with flc pooling are able to disentangle the correct from the incorrect prediction by the prediction confidence. The models provide low-variance and high-confidence in correct predictions and reduced confidence in false predictions across all evaluated samples.

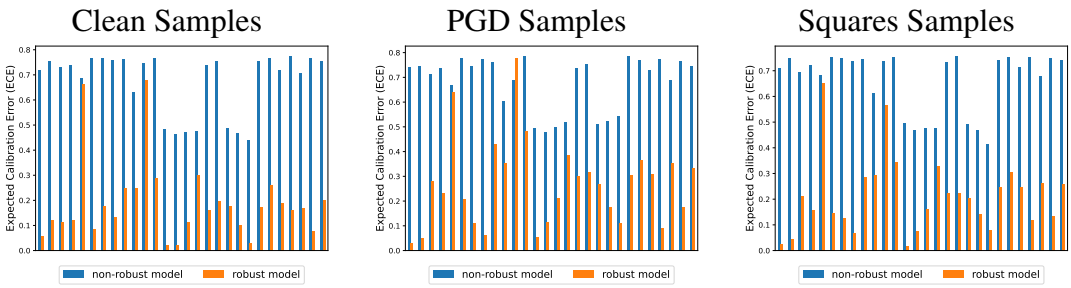

Figure 12: ECE (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR10.

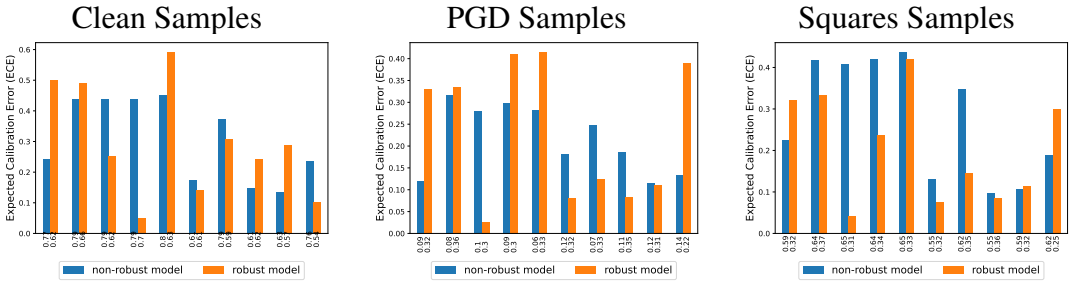

Figure 13: ECE (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR100. The models accuracy are marked for the different samples for each bar.

D Additional Evaluation on ImageNet

Table 3 reports the accuracy evaluation of the robust models as well as the baseline on ImageNet. The accuracy is reported on the clean as well as on the perturbated samples by PGD and Squares with an ϵ of $4/255$.

For completeness, we included the ROC curve on the clean as well as the perturbated samples for the robust models and the baseline on ImagNet in figure 17.

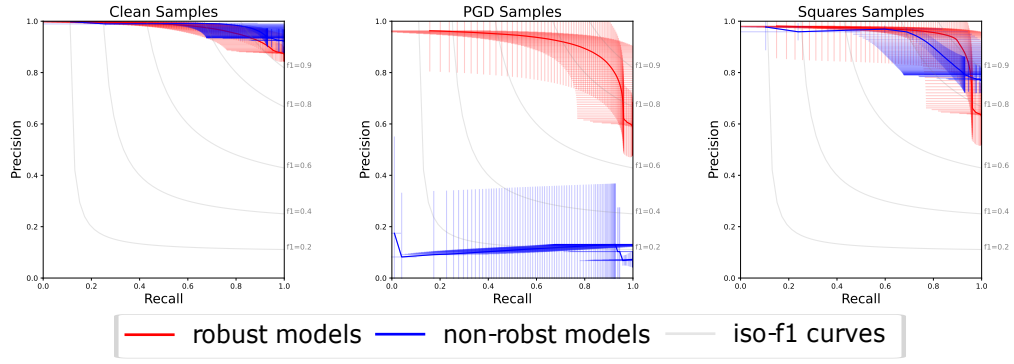

Figure 14: Average precision recall curve for all robust and all non-robust models trained on CIFAR10. Standard deviation is marked by the error bars. For the clean samples, the non-robust models can distinguish slightly better in correct and incorrect predictions based on the confidence of the prediction. The superior of the robust models are visible on the samples created by PGD, the non-robust models are not able to distinguish. However, for the samples created by Squares the classification into correct and incorrect predictions based on the confidence is almost equally possible for robust and non-robust models.

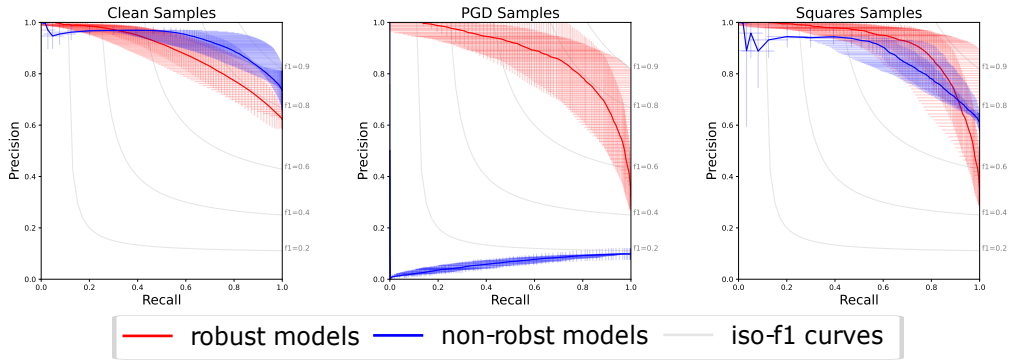

Figure 15: Average precision recall curve for all robust and all non-robust models trained on CIFAR100 for 1000 samples. Standard deviation is marked by the error bars. For the clean samples, the non-robust models can distinguish slightly better in correct and incorrect predictions based on the confidence of the prediction. The superior of the robust models are clearly visible on the samples created by PGD, the non-robust models are not able to distinguish. However, for the samples created by Squares the classification into correct and incorrect predictions based on the confidence is almost equally possible for robust and non-robust models.

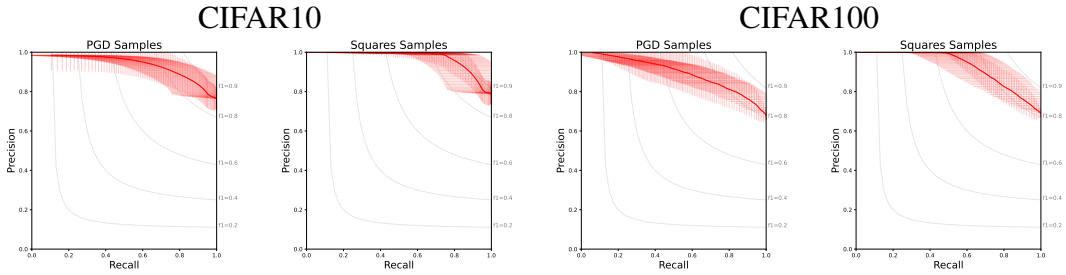

Figure 16: Precision Recall curve between confidence of clean correct samples and perturbated wrong samples on CIFAR10 and CIFAR100. The robust model confidences can be used as threshold for detection of adversarial attacks.

Method	Architecture	Clean Acc ↑	PGD Acc ↑	Squares Acc ↑
Baseline	RN50	**76.13**	0.00	11.48
Engstrom et al. [22]	RN50	62.41	35.47	54.93
Wong et al. [78]	RN50	53.83	29.43	42.26
Salman et al. [63]	RN50	63.87	42.23	56.58
Salman et al. [63]	WRN50-2	68.41	**44.75**	**61.29**
Salman et al. [63]	RN18	52.50	31.92	43.81

Table 3: Clean and robust accuracy against PGD and Squares (higher is better) over 10000 samples.

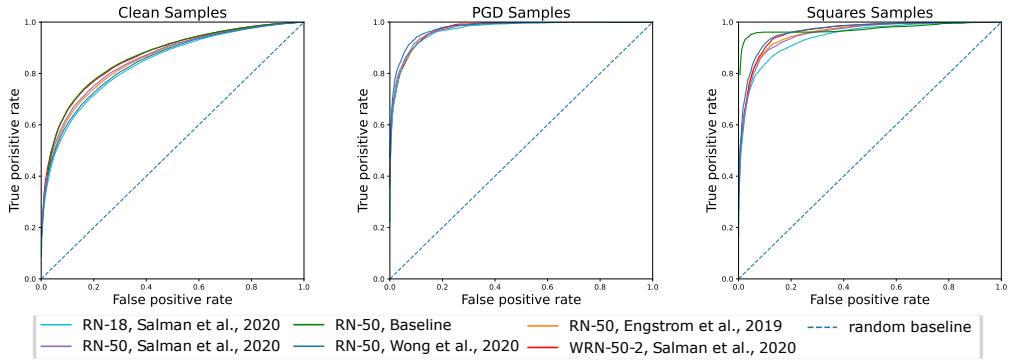

Figure 17: ROC curves for the robust models and the non-robust baseline trained on ImageNet provided on RobustBench [15].

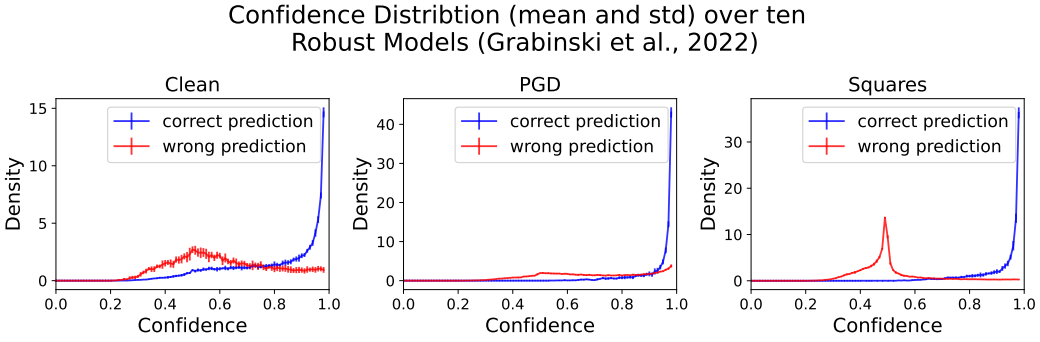

Figure 18: Additional confidence distribution evaluation over ten models (PRN-18) trained on CIFAR10 with flc pooling [27] and AT FGSM [78]. We used 100 bins and present the mean ans standard deviation of the ten different models for each bin.

E Model Overview

The robust checkpoints provided by *RobustBench* [15] are licensed under the MIT Licence. The clean models for ImageNet are provided by *timm* [77] under the Apache 2.0 licence.

Paper	Dataset	Architecture	Adv. Trained Clean Acc.	Adv. Trained Robust Acc.	Norm. Trained Clean Acc.	Norm. Trained Robust Acc.
[3]	cifar10	PreActResNet-18	79.84	43.93	94.51	0.0
[6]	cifar10	WideResNet-28-10	89.69	59.53	95.10	0.0
[64]	cifar10	WideResNet-28-10	88.98	57.14	95.10	0.0
[75]	cifar10	WideResNet-28-10	87.50	56.29	95.10	0.0
[37]	cifar10	WideResNet-28-10	87.11	54.92	95.35	0.0
[61]	cifar10	WideResNet-34-20	85.34	53.42	95.46	0.0
[82]	cifar10	WideResNet-34-10	84.92	53.08	95.26	0.0
[22]	cifar10	ResNet-50	87.03	49.25	94.90	0.0
[11]	cifar10	ResNet-50	86.04	51.56	86.50	0.0
[40]	cifar10	WideResNet-34-10	83.48	53.34	95.26	0.0
[57]	cifar10	WideResNet-34-20	85.14	53.74	76.30	0.0
[78]	cifar10	PreActResNet-18	83.34	43.21	94.25	0.0
[21]	cifar10	WideResNet-28-4	84.36	41.44	94.33	0.0
[81]	cifar10	WideResNet-34-10	87.20	44.83	95.26	0.0
[84]	cifar10	WideResNet-34-10	84.52	53.51	95.26	0.0
[79]	cifar10	WideResNet-28-10	88.25	60.04	95.10	0.0
[79]	cifar10	WideResNet-34-10	85.36	56.17	95.64	0.0
[25]	cifar10	WideResNet-70-16	85.29	57.20	87.91	0.0
[25]	cifar10	WideResNet-70-16	91.10	65.88	87.91	0.0
[25]	cifar10	WideResNet-34-20	85.64	56.86	88.33	0.0
[25]	cifar10	WideResNet-28-10	89.48	62.80	88.20	0.0
[65]	cifar10	WideResNet-34-10	85.85	59.09	95.64	0.0
[65]	cifar10	ResNet-18	84.38	54.43	94.87	0.0
[67]	cifar10	WideResNet-34-10	86.84	50.72	95.26	0.0
[9]	cifar10	WideResNet-34-10	85.32	51.12	95.35	0.0
[16]	cifar10	WideResNet-34-20	88.70	53.57	95.44	0.0
[16]	cifar10	WideResNet-34-10	88.22	52.86	95.26	0.0
[85]	cifar10	WideResNet-28-10	89.36	59.64	95.10	0.0
[59]	cifar10	WideResNet-28-10	87.33	60.75	88.20	0.0
[59]	cifar10	WideResNet-106-16	88.50	64.64	86.92	0.0
[59]	cifar10	WideResNet-70-16	88.54	64.25	87.91	0.0
[59]	cifar10	WideResNet-70-16	92.23	66.58	87.91	0.0
[68]	cifar10	WideResNet-28-10	89.46	59.66	95.10	0.0
[68]	cifar10	WideResNet-34-15	86.53	60.41	95.50	0.0
[59]	cifar10	PreActResNet-18	83.53	56.66	89.01	0.0
[58]	cifar10	PreActResNet-18	89.02	57.67	89.01	0.0
[58]	cifar10	PreActResNet-18	86.86	57.09	89.01	0.0
[58]	cifar10	WideResNet-34-10	91.47	62.83	88.67	0.0
[58]	cifar10	WideResNet-28-10	88.16	60.97	88.20	0.0
[39]	cifar10	WideResNet-34-R	90.56	61.56	95.60	0.0
[39]	cifar10	WideResNet-34-R	91.23	62.54	95.60	0.0
[1]	cifar10	ResNet-18	80.24	51.06	94.87	0.0
[1]	cifar10	WideResNet-34-10	85.32	58.04	95.26	0.0
[26]	cifar10	WideResNet-70-16	88.74	66.11	87.91	0.0
[17]	cifar10	WideResNet-28-10-PSSiLU	87.02	61.55	85.53	0.0
[26]	cifar10	WideResNet-28-10	87.50	63.44	88.20	0.0

Continued on next page

Paper	Dataset	Architecture	Adv. Trained Clean Acc.	Adv. Trained Robust Acc.	Norm. Trained Clean Acc.	Norm. Trained Robust Acc.
[26]	cifar10	PreActResNet-18	87.35	58.63	89.01	0.0
[8]	cifar10	WideResNet-34-10	85.21	56.94	95.64	0.0
[8]	cifar10	WideResNet-34-20	86.03	57.71	95.29	0.0
[25]	cifar100	WideResNet-70-16	60.86	30.03	60.56	0.0
[25]	cifar100	WideResNet-70-16	69.15	36.88	60.56	0.0
[16]	cifar100	WideResNet-34-20	62.55	30.20	80.46	0.0
[16]	cifar100	WideResNet-34-10	70.25	27.16	79.11	0.0
[16]	cifar100	WideResNet-34-10	60.64	29.33	79.11	0.0
[9]	cifar100	WideResNet-34-10	62.15	26.94	78.75	0.0
[79]	cifar100	WideResNet-34-10	60.38	28.86	78.79	0.0
[67]	cifar100	WideResNet-34-10	62.82	24.57	79.11	0.0
[37]	cifar100	WideResNet-28-10	59.23	28.42	79.16	0.0
[61]	cifar100	PreActResNet-18	53.83	18.95	76.18	0.0
[59]	cifar100	WideResNet-70-16	63.56	34.64	60.56	0.0
[59]	cifar100	WideResNet-28-10	62.41	32.06	61.46	0.0
[58]	cifar100	PreActResNet-18	56.87	28.50	63.45	0.0
[58]	cifar100	PreActResNet-18	61.50	28.88	63.45	0.0
[1]	cifar100	PreActResNet-18	62.02	27.14	76.66	0.0
[1]	cifar100	WideResNet-34-10	65.73	30.35	79.11	0.0
[8]	cifar100	WideResNet-34-10	64.07	30.59	79.11	0.0
[78]	imagenet	ResNet-50	55.62	26.24	80.37	0.0
[22]	imagenet	ResNet-50	62.56	29.22	80.37	0.0
[63]	imagenet	ResNet-50	64.02	34.96	80.37	0.0
[63]	imagenet	ResNet-18	52.92	25.32	69.74	0.0
[63]	imagenet	WideResNet-50-2	68.46	38.14	81.45	0.0