Supplementary Material

1 Adversarial Perturbations and Spectral Distribution

Figures 1–4 show the natural images, adversarial images, and the spectral distribution (low frequency in the center) of the perturbations across the datasets. x denotes the natural images, δ_{nm}, δ_{lm} and δ_{rm} denote the PGD-20 attack perturbations generated according to the natural, L- and robust models, respectively. *FFT* denotes the Fast Fourier Transform. Jet color map is used to highlight perturbations for clear visualization. For the **natural models**, the perturbations are a jumble of noise points within the picture and have large magnitudes in the high-frequency region. The perturbations for the **adversarial models** are significantly more ordered and mainly concentrated in the low-frequency region. These visualizations prove that the adversarial perturbation is **not** a simple high-frequency phenomenon and is **model- and dataset-dependent**.

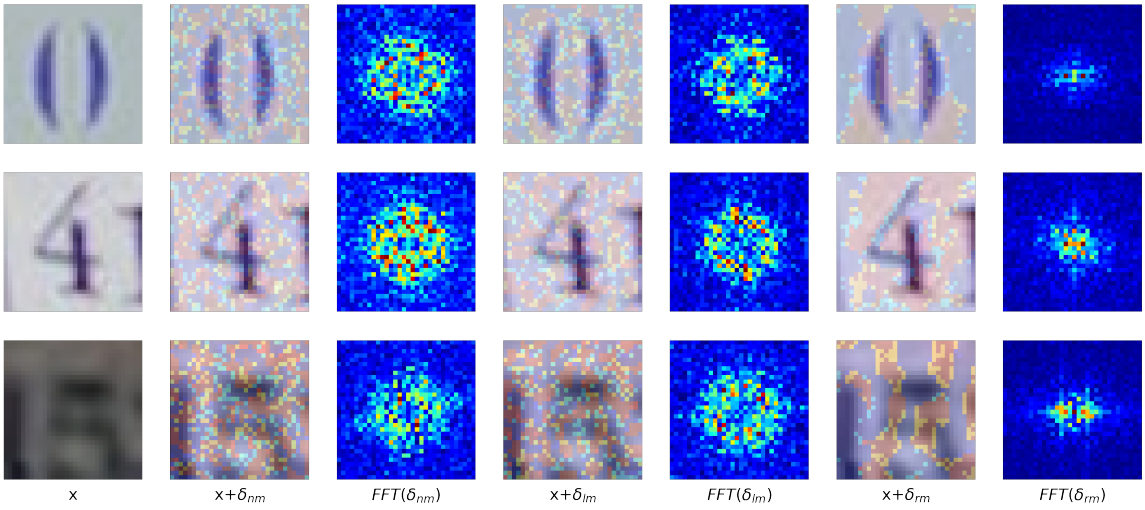

Figure 1: Visualization of the natural and perturbed images on SVHN.

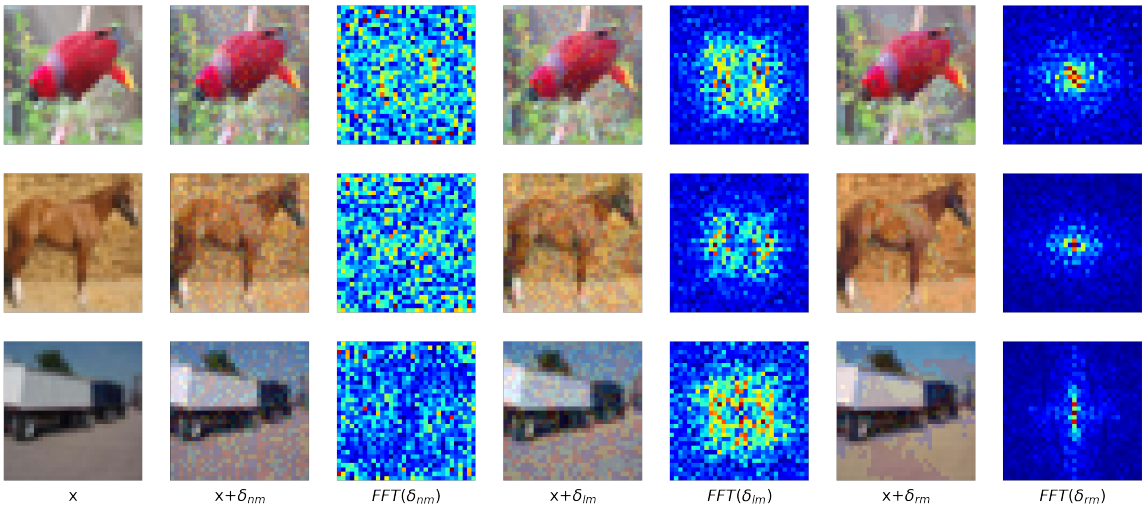

Figure 2: Visualization of the natural and perturbed images on CIFAR-10.

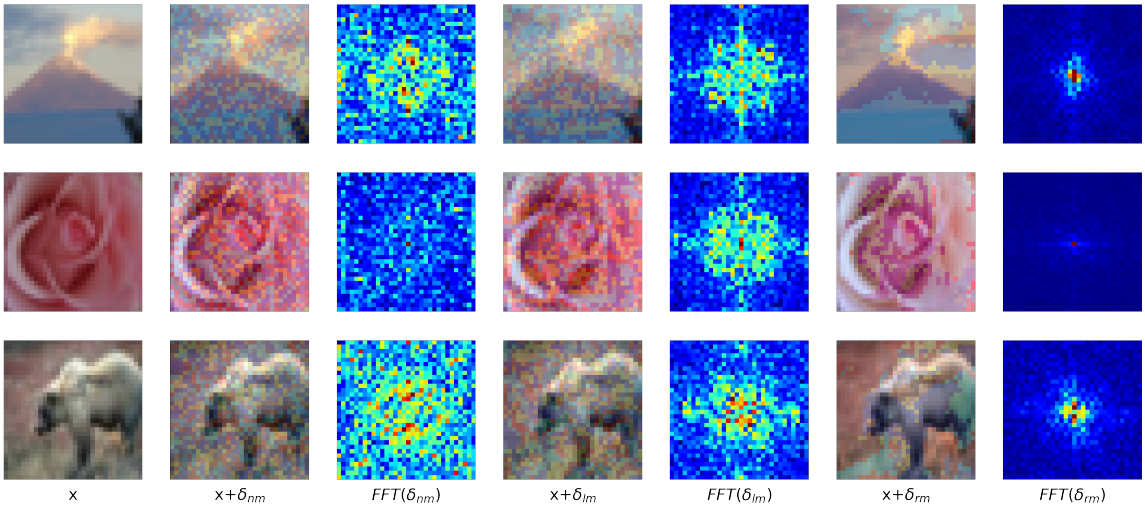

Figure 3: Visualization of the natural and perturbed images on CIFAR-100.

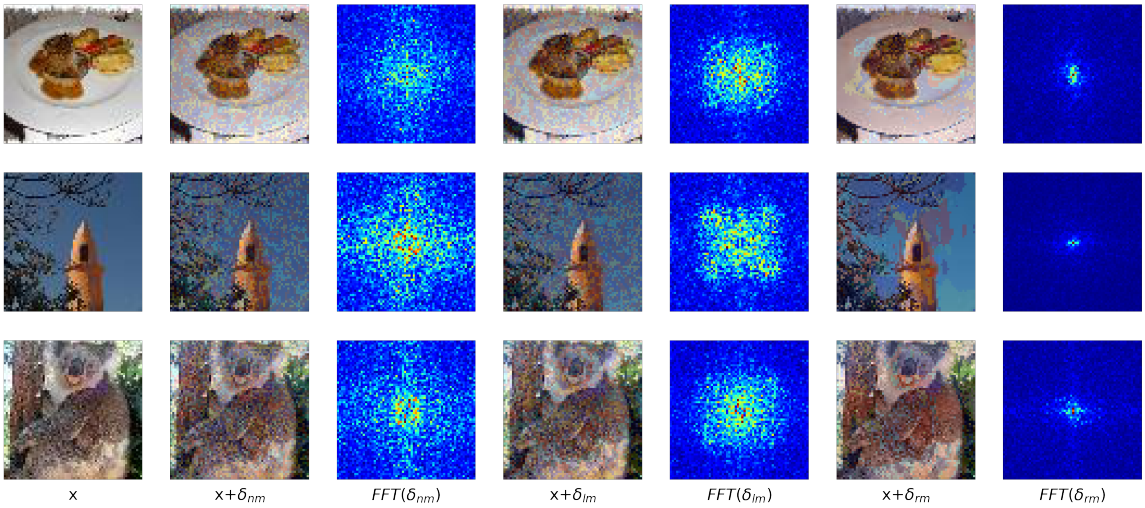

Figure 4: Visualization of the natural and perturbed images on Tiny ImageNet.

2 Detailed Settings for Popular Defenses

TRADES: It decomposes the robust error as the sum of the natural error and the boundary error and encourages the algorithm to push the decision boundary away from the data to improve the robust accuracy. The overall loss function is shown as follows:

$$\mathcal{L}_{AT} = \mathbf{CE}(f(x), y) + \lambda \cdot \mathbf{KL}(f(x) \| f(x+\delta)) \tag{2.1}$$

CE denotes the Cross-Entropy loss, **KL** denotes the Kullback-Leibler divergence generated by PGD-10, δ denotes the adversarial perturbations, $f(x)$ denotes the probability predicted by the model, y denotes the true label, λ is the coefficient to balance the CE and KL loss. Following the default setting in TRADES, we adopt SGD with momentum 0.9, weight decay 2×10^{-4}, and batch size 128. The model is trained for 100 (30)[1]. epochs on one 3090 GPU. The initial learning rate is 0.1 (0.01), which decays to one-tenth at 75th (15th) and 90th (25th) epochs, respectively. The λ in Eqn. 2.2 is set to 6.

MART: Based on standard AT, it explicitly differentiates the misclassified and correctly classified examples during the training and adds a misclassification-aware regularization to the standard adversarial risk to achieve better robustness. The overall loss function is shown as follows:

$$\mathcal{L}_{AT} = \mathbf{BCE}(f(x+\delta), y) + \lambda \cdot \mathbf{KL}(f(x) \| f(x+\delta)) \cdot (1 - f(x)) \tag{2.2}$$

BCE denotes the binary Cross-Entropy loss. Following the default setting in MART, we adopt SGD with momentum 0.9, weight decay 2×10^{-4}, and batch size 128. The model is trained for 100 (30) epochs on one 3090 GPU. The initial learning rate is 0.1 (0.01), which decays to one-tenth at 75th (15th) and 90th (25th) epochs, respectively. The λ in Eqn. 2.2 is set to 6.

AWP: It identifies the connection between the weight loss landscape and the robust generalization gap, proposes adversarial weight perturbation to directly make the weight loss landscape flat, and develops a double perturbation (adversarially perturbing both inputs and weights) mechanism in the AT framework. Following the default setting in AWP, we adopt SGD with momentum 0.9, weight decay 5×10^{-4}, and batch size 128. The model is trained for 200 epochs on two V100 GPUs. The initial learning rate is 0.1 (0.01), which decays to one-tenth at 100th and 150th epochs, respectively.

[1]The numbers in brackets are the hyperparameters for SVHN.

3 More Comparisons with FR

Experimental setup: For a fair comparison, all experiments adopt the same data augmentation method: 4-pixel padding with 32×32 random crops (not for SVHN and Tiny ImageNet) and random horizontal flip (not for SVHN). All natural images are normalized to $[0, 1]$. The Frequency Regularization (FR) coefficient is set to 0.1 for SVHN and CIFAR datasets, and 0.05 for Tiny ImageNet. The training set was randomly divided into the training set and the validation set according to the ratio of 9:1. We select the model with the highest robustness against PGD-20 attacks on the validation set for further evaluation against other popular attacks.

Effectiveness across the datasets and methods: We provide the thorough performance comparison of AT+FR and other methods on ResNet18 in Table 1-4. Since the FR and FR/WA are plug-and-play blocks, we also apply them to popular techniques to prove their effectiveness. Experimental results demonstrate that FR can be plugged into these popular methods to further improve robustness. Besides, FR/WA can maintain a similar standard accuracy as AT while improving robust accuracy. The improvement is non-trivial since some papers have claimed a trade-off between the standard and robust accuracy.

Table 1: Top-1 accuracy(%) of the ResNet18 model on the SVHN. Bold numbers indicate the best.

Method	Clean	PGD-20	PGD-50	C&W	AA
AT	**90.88**	53.28	52.26	50.62	47.57
AT + FR	90.43	56.87	56.29	**51.89**	**49.46**
AT + FR/WA	90.49	**56.95**	**56.35**	51.76	49.36
TRADES	**89.04**	55.71	55.48	51.48	50.03
TRADES + FR	87.23	56.96	56.73	**52.64**	**51.26**
TRADES + FR/WA	88.76	**57.01**	**56.76**	52.53	51.18
MART	**88.67**	56.78	56.40	50.03	47.75
MART + FR	86.42	**57.87**	**57.52**	**50.65**	**48.74**
MART + FR/WA	88.53	57.58	57.26	50.42	48.58
AWP	**89.12**	53.85	53.23	49.84	47.05
AWP + FR	86.63	**55.64**	**55.33**	**50.83**	**49.53**
AWP + FR/WA	88.73	55.32	55.11	50.68	49.37

Table 2: Top-1 accuracy(%) of the ResNet18 model on the CIFAR-10. Bold numbers indicate the best.

Method	Clean	PGD-20	PGD-50	C&W	AA
AT	**81.98**	51.69	51.46	50.44	48.19
AT + FR	80.04	**55.39**	**55.13**	51.61	50.02
AT + FR/WA	81.74	55.12	54.88	**52.21**	**50.16**
TRADES	**81.83**	53.41	53.23	50.92	49.84
TRADES + FR	80.11	**54.45**	**54.24**	**51.51**	**50.52**
TRADES + FR/WA	81.76	54.14	54.06	51.31	50.38
MART	**81.01**	54.58	54.47	50.01	48.10
MART + FR	79.03	**55.17**	**54.90**	**50.98**	**49.22**
MART + FR/WA	80.80	55.01	54.78	50.12	48.78
AWP	**81.06**	55.36	55.27	51.98	50.37
AWP + FR	79.03	**57.07**	**57.01**	**52.16**	**50.80**
AWP + FR/WA	80.87	56.81	56.82	52.14	50.61

Table 3: Top-1 accuracy(%) of the ResNet18 model on the CIFAR-100. Bold numbers indicate the best.

Method	Clean	PGD-20	PGD-50	C&W	AA
AT	**54.18**	27.81	27.49	25.82	23.56
AT + FR	49.23	31.27	31.20	27.60	**26.09**
AT + FR/WA	53.66	**31.49**	**31.36**	**28.24**	26.06
TRADES	56.24	28.48	28.40	24.71	23.77
TRADES + FR	55.57	30.08	29.95	25.93	25.05
TRADES + FR/WA	**56.59**	**30.22**	**30.20**	**26.81**	**25.37**
MART	**51.23**	29.66	29.55	25.88	24.27
MART + FR	49.33	31.03	30.87	26.78	25.07
MART + FR/WA	50.72	**31.75**	**31.68**	**27.32**	**25.46**
AWP	54.71	30.88	30.69	27.87	25.74
AWP + FR	48.92	**31.90**	**31.73**	28.02	26.10
AWP + FR/WA	**55.78**	31.75	31.60	**28.84**	**26.64**

Table 4: Top-1 accuracy(%) of the ResNet18 model on the Tiny ImageNet. Bold numbers indicate the best.

Method	Clean	PGD-20	PGD-50	C&W	AA
AT	**46.64**	23.33	23.18	20.44	18.34
AT + FR	42.92	25.32	25.25	21.35	**19.71**
AT + FR/WA	46.55	**25.64**	**25.47**	**21.90**	19.64
TRADES	48.33	22.77	22.71	18.89	17.95
TRADES + FR	46.61	23.79	23.72	19.97	18.99
TRADES + FR/WA	**48.94**	**24.81**	**24.73**	**20.21**	**19.34**
MART	**45.39**	24.17	24.09	20.07	18.49
MART + FR	43.35	25.35	25.31	20.63	19.31
MART + FR/WA	44.79	**25.83**	**25.80**	**21.53**	**19.44**
AWP	**46.49**	24.76	24.59	21.04	19.11
AWP + FR	44.93	**24.98**	**24.90**	**21.54**	**19.52**
AWP + FR/WA	46.32	24.86	24.77	21.32	19.38

4 Large Architecture

To prove the effectiveness of the FR and FR/WA blocks on large models, we trained the ResNet152 and Wide ResNet-34-15 on the CIFAR-10 dataset. The results are shown in Table 5. For reference, the ResNet18 and Wide ResNet-34-10 have a parameter count of 11.174M and 46.160M, respectively. For large models, FR and FR/WA can still improve the model's robustness against multiple attacks relative to the standard AT, proving that the proposed methods are suitable for large models.

Table 5: Top-1 robust accuracy(%) of various models on the CIFAR-10. #number indicates the parameters. Bold numbers indicate the best.

model	Method	Clean	PGD-20	PGD-50	C&W	AA
ResNet152 # 58,156,618	AT	84.60	54.62	54.44	52.86	50.87
	AT + FR	81.11	**57.76**	**57.70**	54.74	53.32
	AT + FR/WA	**84.62**	57.55	57.50	**55.63**	**54.10**
Wide ResNet-34-15 # 103,819,674	AT	**86.66**	55.79	55.49	54.58	52.76
	AT + FR	84.35	**58.33**	**58.03**	**55.75**	**54.12**
	AT + FR/WA	86.18	57.97	57.78	55.25	53.50

5 Failure Cases

We refer to the successfully evaluated images as success cases and incorrectly classified ones as failure cases. Taking the adversarially trained ResNet18 with FR on CIFAR-10 as an example, we visualize randomly selected images and the average Fourier spectra of all perturbations in Figure 5. The perturbations in failure cases are more concentrated in the low-frequency center than in success cases. This indicates that it is essential to improve the resistance of the model to ultra-low-frequency perturbations to further improve the robustness.

Figure 5: Visualization of the success and failure images on CIFAR-10. The last column is the average Fourier spectra of the PGD-20 attack perturbations with low frequency in the center.