Supplementary Materials: Feature Separation and Recalibration for Adversarial Robustness

Woo Jae Kim Yoonki Cho Junsik Jung Sung-Eui Yoon
Korea Advanced Institute of Science and Technology (KAIST)

ResNet-18	CIFAR-100					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	**59.25**	28.80	24.39	23.43	23.92	22.46
AT + FSR	58.23	**29.58**	**25.33**	**24.30**	**24.54**	**22.95**
TRADES	**61.87**	30.77	26.37	25.76	24.08	23.45
TRADES + FSR	57.27	**31.66**	**27.70**	**27.27**	**24.82**	**24.40**
MART	**57.13**	31.32	27.40	26.80	25.24	24.42
MART + FSR	56.51	**32.08**	**27.90**	**27.28**	**25.91**	**24.98**

Table S1. Robustness (accuracy (%)) of adversarial training strategies (AT, TRADES, MART) with (+ FSR) and without our FSR module against diverse white-box attacks on ResNet-18 and on CIFAR-100 dataset. Better results are marked in **bold**.

ResNet-18	Tiny ImageNet					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	51.13	22.54	18.69	17.87	17.83	16.34
AT + FSR	**51.77**	**24.19**	**20.95**	**20.06**	**19.32**	**18.02**
TRADES	**50.41**	23.79	21.16	20.72	17.24	17.02
TRADES + FSR	49.53	**24.87**	**23.22**	**23.09**	**19.22**	**19.04**
MART	**46.21**	23.84	21.75	21.35	18.34	17.71
MART + FSR	46.02	**26.02**	**24.05**	**23.82**	**20.63**	**20.24**

Table S2. Robustness (accuracy (%)) of adversarial training strategies (AT, TRADES, MART) with (+ FSR) and without our FSR module against diverse white-box attacks on ResNet-18 and on Tiny ImageNet dataset. Better results are marked in **bold**.

1. Additional Robustness Evaluation

In this section, we report the robustness of our FSR on additional datasets (CIFAR-100 [6], Tiny ImageNet [4]) and model (WideResNet-34-10 [8]).

Experiments on Other Datasets. Table S1 shows the robustness improvements when our FSR module is applied on AT, TRADES, and MART in CIFAR-100 dataset. While the performance improvements are not as large as in CIFAR-10 and SVHN, applying our FSR module consistently improves the model robustness of all three adversarial training techniques, showing that our method is still effective on more challenging datasets. We noted that the reason for limited accuracy gain on CIFAR-100 is actually due to its low-resolution data not providing sufficient information for learning the inter-class relationship among cues relevant to various similar classes (e.g., boy and man) [3].

WideResNet-34-10	CIFAR-10					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	**87.49**	59.47	50.72	48.75	50.42	48.52
AT + FSR	87.02	**61.40**	**53.78**	**52.04**	**52.35**	**50.36**
TRADES	86.06	60.78	51.77	49.66	51.34	49.27
TRADES + FSR	**86.88**	**62.97**	**54.37**	**51.98**	**53.19**	**51.34**
MART	85.81	61.22	52.49	49.88	49.67	48.81
MART + FSR	**86.21**	**62.61**	**54.23**	**52.00**	**51.25**	**50.10**

Table S3. Robustness (accuracy (%)) of adversarial training strategies (AT, TRADES, MART) with (+ FSR) and without our FSR module against diverse white-box attacks on WideResNet-34-10 and on CIFAR-10 dataset. Better results are marked in **bold**.

Thus, we also evaluate our method on a more challening Tiny ImageNet dataset with fine-grained classes and higher-resolution images. As shown by the results in Table S2, we observed 2.08% improvement on average for Ensemble robustness compared to vanilla methods, which is significantly higher than that of CIFAR-100 (0.67%, Table S1) and on par with CIFAR-10 (2.20%, Table 1) and SVHN (2.30%, Table 2). This shows that our FSR module is also effective on larger, more complex models and datasets and is not limited by the over-parameterization of the model.

Experiments on Other Model. In addition to ResNet-18 and VGG16 in the main paper, we also evaluate our FSR module on WideResNet-34-10. As shown in Table S3, our FSR module leads to consistent robustness improvement on WideResNet-34-10.

2. Additional Ablation Studies

Position of FSR module. Table S4 reports the model robustness when our FSR module is inserted to different layers of ResNet-18. As shown in the table, inserting our FSR module after `Block4` of the model shows the best model robustness under attacks. This is because the model learns features that are more related to the global semantic information of the image and the final class prediction in the deeper layers, while it learns more low-level features with less semantic information in shallower layers [2]. Recalibrating the non-robust activations in the deeper layers that are more related to the final predictions is more effective at

	No attack	FGSM	PGD-20	PGD-100	C&W	Ensemble
Block1	**84.58**	56.41	48.29	46.28	46.96	44.89
Block2	83.76	56.34	48.86	47.03	47.32	45.28
Block3	82.60	56.62	50.43	49.11	47.84	46.33
Block4	81.46	**58.07**	**52.47**	**51.02**	**49.44**	**48.34**
Block3 + Block4	82.18	56.93	50.72	49.32	48.63	46.91

Table S4. Comparison of accuracy (%) as we insert our FSR module after different layers of ResNet-18.

	No attack	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	85.02	56.21	48.22	46.37	47.38	45.51
Uniform	**85.16**	58.05	50.87	48.91	**49.99**	47.90
Entropy max.	84.69	**58.35**	50.66	48.93	49.90	47.88
Avg. targeted loss	84.50	57.98	50.41	48.55	49.80	47.42
Mispredicted (Ours)	81.46	58.07	**52.47**	**51.02**	49.44	**48.34**

Table S5. Comparison of accuracy (%) for different design choices of the separation loss $\mathcal{L}_{sep}$ (Eq. 3).

boosting the model robustness.

Design Choice of $\mathcal{L}_{sep}$. As explained in Sec. 3.1 of the main paper, in order to disentangle the non-robust activations through the separation loss $\mathcal{L}_{sep}$ (Eq. 3 of the main paper):

$$\mathcal{L}_{sep} = -\sum_{i=1}^{N}(y_i \cdot \log(p_i^+) + y_i' \cdot \log(p_i^-)), \qquad (1)$$

we minimize the cross entropy loss of the prediction score with respect to y', which we define as the label corresponding to the wrong class with the highest prediction score. In Table S5, we report the comparison of robustness as we employ different schemes for such disentanglement. "Uniform" represents replacing y' with a uniform vector implemented through label smoothing, "Entropy max." represents maximizing the entropy of the output prediction p^- on the non-robust feature, "Avg. targeted loss" represents the average of cross-entropy loss with respect to all class labels except for the ground truth class, and "Mispredicted" represents our original design. All four schemes lead to meaningful improvement compared to the vanilla AT method, as they guide the Separation Net to learn low robustness scores on feature units that are responsible for predictions other than the ground truth class. Still, our design of using the mispredicted class output achieves the highest robustness under all attacks. This implies that through this scheme, the Separation Net learns to assign low robustness scores to the most harmful feature units that lead to the most probable model mistake and thus improves the feature robustness by the largest margin.

Effects of Gumbel Softmax. We verify the effects of applying Gumbel softmax to generate a differentiable soft mask m that divides the input feature map into the robust activations and the non-robust activations. We compare the robustness upon replacing m with a binary mask b (Sec. 3.1 of the main paper) that divides the activations in a discrete manner. We implement the binary mask b by first applying a sigmoid normalization function to the robustness map r generated by the Separation Net and setting all values less than 0.5 to 0 and all values greater than or equal to 0.5 to 1. In other words, for an i-th unit of the robustness map r, we set b_i as follows:

$$b_i = \begin{cases} 0, & \text{if } \sigma(r)_i < t \\ 1, & \text{if } \sigma(r)_i \geq t, \end{cases} \qquad (2)$$

where $t = 0.5$, and $\sigma(\cdot)$ is the sigmoid normalization function.

	FGSM	PGD-20	PGD-100	C&W	Ensemble	AutoAttack
Binary	55.78	49.21	47.79	48.74	46.91	44.26
Gumbel	**58.07**	**52.47**	**51.02**	**49.44**	**48.34**	**46.41**

Table S6. Comparison of accuracy (%) on using mask generated by discrete binary sampling or through Gumbel softmax.

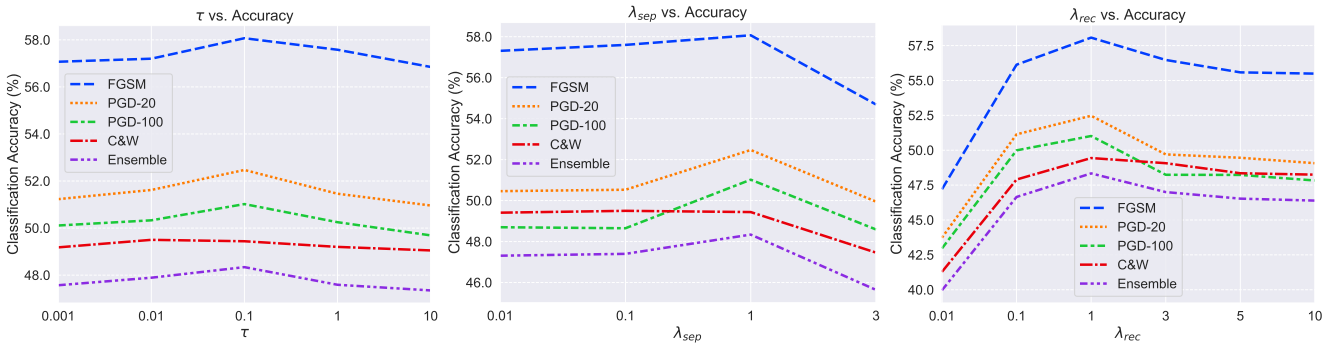
	FGSM	PGD-20	PGD-100	C&W	Ensemble
Greedy	57.75	49.48	47.59	48.36	46.42
Random	56.60	50.04	48.46	49.08	46.77
w/o Separation	57.51	50.71	48.98	49.32	47.60
w/ Separation (Ours)	**58.07**	**52.47**	**51.02**	**49.44**	**48.34**

Table S7. Comparison of accuracy (%) as we replace the Separation Net with different strategies.

In Table S6, we show the comparison of robustness of our method upon using either b (Binary) or m (Gumbel). Using the differentiable mask m through the Gumbel softmax leads to higher robustness against all white-box attacks and especially against the AutoAttack than using the binary mask b. Using the Gumbel softmax allows us to learn the mask to better capture the feature robustness, and it also prevents gradient masking, thus showing higher robustness against AutoAttack.

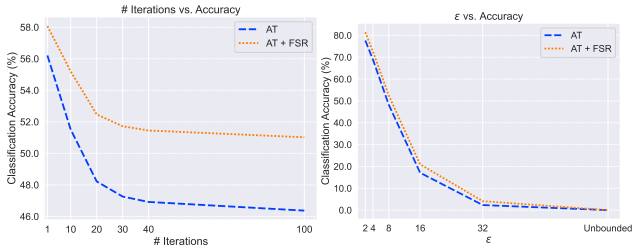
Experiments on Effectiveness of the Separation Net. In order to verify whether our Separation Net is learning appropriate robustness scores for each feature activaiton, we tried replacing the output mask m from the Separation Net (Eq. 2) with different strategies. We tested random selection and a greedy method of recalibrating the lowest activations, both of which would recalibrate feature activations unaware of their robustness. Table S7 shows that both strategies significantly lag behind our method without Separation, which is equivalent to recalibrating all activations (refer to Table 6). This is because they do not fully recapture the discriminative cues underlying in non-robust activations. Our method with Separation leads to the highest robustness, showing that FSR well identifies the non-robust activations and recaptures discriminative cues from them.

Hyperparameter Study. We also compare the robustness



(a) τ (b) λ_{sep} (c) λ_{rec}

Figure S1. Analysis on the robustness with various values of hyperparameters used in FSR module. (a) Study on τ that controls the temperature on Gumbel softmax. (b) Study on λ_{sep} that controls the weight on the separation loss $\mathcal{L}_{sep}$. (c) Study on λ_{rec} that controls the weight on the recalibration loss $\mathcal{L}_{rec}$.



(a) Varying # iterations (b) Varying ϵ

Figure S2. Analysis on obfuscated gradients. (a) Change in model robustness as we vary the number of iterations in PGD attack. (b) Change in model robustness as we vary the perturbation bound ϵ.

as we vary the temperature τ (Eq. 2) that controls how "discrete" the mask is. For low temperature values, the output mask becomes more discrete (*i.e.*, most values are close to either 0 or 1), and for high temperature values, it becomes more uniform (*i.e.*, most values are far away from 0 or 1) [5]. As shown in Fig. S1a, we achieve the highest robustness when $\tau = 0.1$. From this observation, we can see that too small τ will degenerate the Gumbel softmax into binary sampling and make the mask become a binary mask, which could result in no gradients or improper training [7]. In contrast, too large τ will make the mask become more uniformly distributed and reduce the gap between the mask values applied on robust or non-robust activations, thus making our goal of disentanglement less feasible.

In Fig. S1b and Fig. S1c, we visualize the trends of model robustness as we vary the weights on our proposed loss functions $\mathcal{L}_{sep}$ (Eq. 3) and $\mathcal{L}_{rec}$ (Eq. 4). Higher value of λ_{sep} generally improves robustness under all attacks with the best performance achieved when $\lambda_{sep} = 1$, showing that our proposed objectives help the model learn more robust feature representations. Similar trends can also be observed for λ_{rec}; higher value of λ_{rec} generally improves robustness with the best performance achieved when $\lambda_{rec} = 1$. Setting λ_{sep} and λ_{rec} to be too high, however, tends to degrade robustness. This is because of the trade-off between the vanilla classification loss $\mathcal{L}_{cls}$ on the final classifier layer and the two auxiliary loss. As we focus more on the objectives on the auxiliary layer, the two auxiliary losses may deviate the model from learning the classification task based on $\mathcal{L}_{cls}$.

3. Analysis on Obfuscated Gradients

In this section, we verify that the robustness of our method is not a result of obfuscating gradients. We test our method under the following criteria [1] to demonstrate that our method does not obfuscate gradients:

(i) White-box attacks are stronger than black-box attacks,

(ii) Robustness decreases with the increased number of iterations in gradient-based attacks,

(iii) Robustness decreases with increased perturbation bound ϵ, and unbounded attacks achieve 100% attack success rate.

Tables 1 and 3 of the main paper show the robustness of our method under both white-box and black-box attacks when applied to ResNet-18 on the CIFAR-10 dataset. Comparing the two tables, we can observe that the strongest black-box attacks (*e.g.*, DI-FGSM and $\mathcal{N}$Attack) are still weaker than white-box attacks (*e.g.*, C&W), meeting the requirement (i). Fig. S2a shows robustness of our method and vanilla PGD adversarial training under PGD attacks with various number of iterations. The robustness does indeed decrease with increasing number of iterations, meeting the requirement (ii). Fig. S2b shows robustness of the two methods under PGD attacks with various perturbation bounds ϵ under ℓ_∞-norm. Similarly, the robustness decreases with increasing ϵ, and it reaches 0% accuracy under unbounded attacks, thus meeting the requirement (iii).

Acknowledgement Prof. Sung-Eui Yoon is a corresponding author. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208506).

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 3

[2] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *ICLR*, 2021. 1

[3] Dingding Cai, Ke Chen, Yanlin Qian, and Joni-Kristian Kämäräinen. Convolutional low-resolution fine-grained classification. In *Pattern Recognition Letters*, 2019. 1

[4] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *ICCV*, 2021. 1

[5] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 3

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[7] Zhenda Xie, Zheng Zhang, Xizhou Zhu, Gao Huang, and Stephen Lin. Spatially adaptive inference with stochastic feature sampling and interpolation. In *ECCV*, 2020. 3

[8] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 1