A. Details on Visualizing Channel-wise Activations

A.1. Non-robust CNNs vs. Robustified CNNs

We train ResNet-18 models to perform the classification task on the CIFAR10 dataset. The models are trained normally and adversarially. We use adversarial data generated by PGD-10 attack ($\epsilon = 8/255$, step size $\epsilon/4$, and random initialization) for adversarial training.

The ResNet-18 network consists of one convolutional layer, eight residual blocks, and one linear fully-connected (FC) layer connected successively. Each residual block contains two convolutional layers for the residual mapping. We visualize the features of the penultimate layer (the output of the eighth residual block) and the weights of the last linear layer in Figure 1. Specifically, the weights of the last FC layer for a certain class are sorted and plotted in descending order. We process the penultimate layer's features with the global average pooling operation to obtain the channel-wise activations. For a certain class, we calculate each channel's mean activation magnitude over all the test samples in this category. We normalize the mean channel-wise activations by dividing them by their absolute maximum. The mean channel-wise activations are plotted according to the indices of the sorted weights. We also record the activated frequency of each channel. Here, the channel is regarded to be activated if its activation magnitude is larger than a threshold (1% of the maximum of all channels' activations).

A.2. Channel-wise Activations of CIFS-modifed CNNs

We train CIFS-modified CNNs normally and adversarially by using the adaptive loss in Equation (3). We use the PGD-10 attack to generate adversarial data. We illustrate the channels' activations of CIFS-modified CNNs in Figure 3. The implementation details are same as those in Appendix A.1.

In Figure 3, we show the channels' activations of data in class "airplane". Here, we plot the channels activations of data in other classes. From Figure A.2, we see that CIFS indeed suppresses negatively-relevant (NR) channels and promotes the positively-relevant (PR) ones.

Besides, we also observe that CIFS ameliorates the class-wise imbalance of robustness under AT. In Figure A.1, we can see that, for the data in class "cat" and class "deer", the robust accuracies of the vanilla ResNet-18 model are 16.70% and 25.50%. Modifying the vanilla model with CIFS-softmax, we can improve the robust accuracies by 5.6 and 3.3 percentage points, respectively.

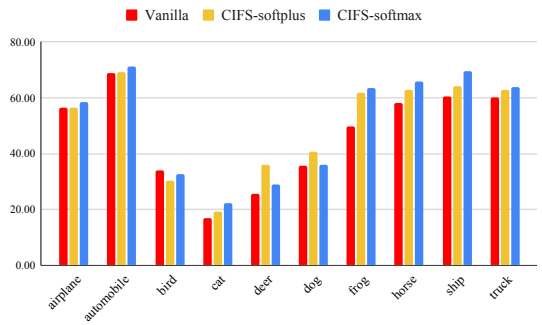


Figure A.1. Robust accuracies (%) of PGD-20 adversarial data for various classes of the CIFAR10 dataset.

B. Robustness Evaluation on CIFAR10

B.1. Robustness Enhancement of CIFS under AT

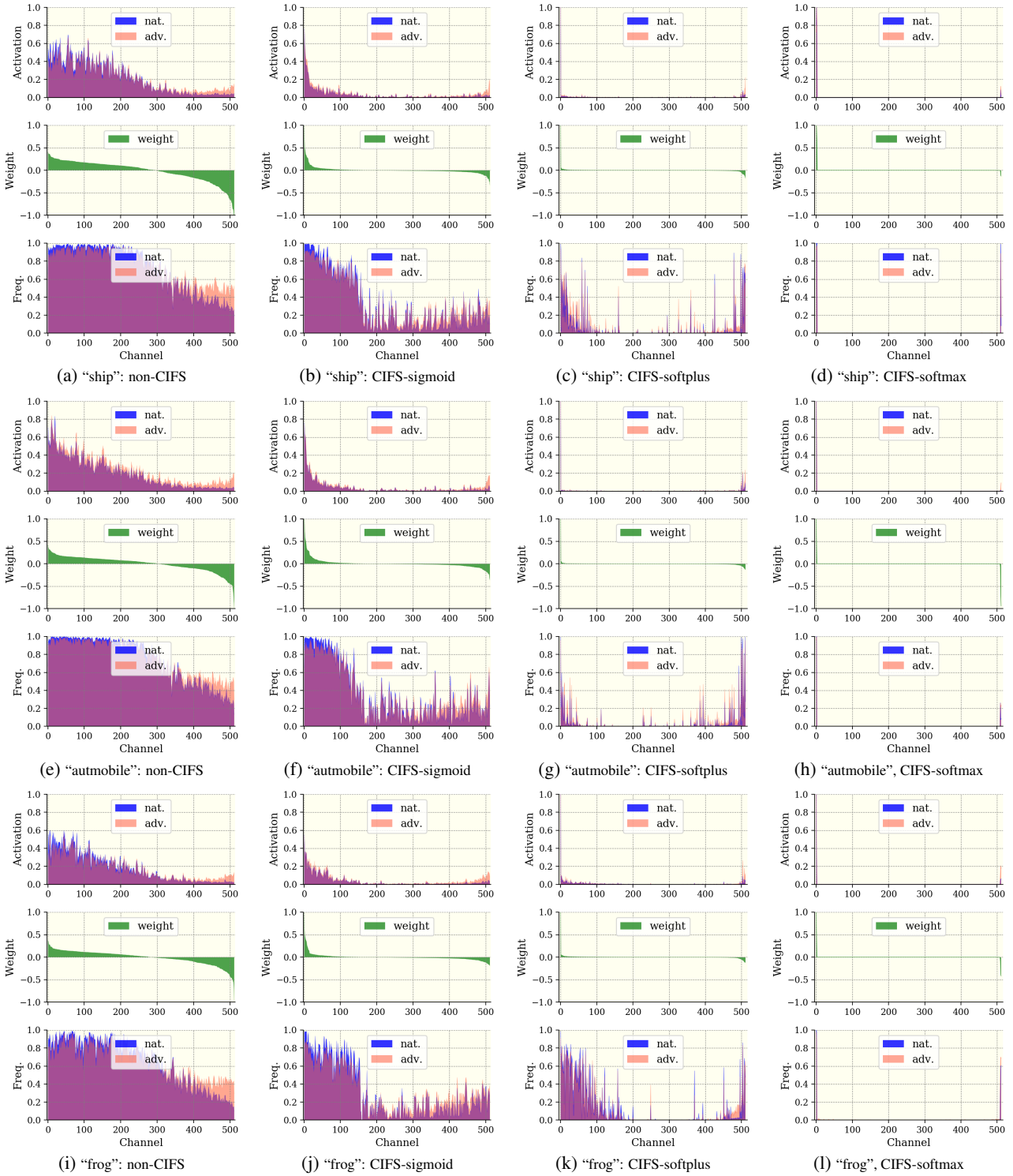
Training and Evaluation details: On the CIFAR10 dataset, we train ResNet-18 and WRN-28-10 models with PGD-10 adversarial examples ($\epsilon = 8/255$, step size $\epsilon/4$ with random initialization). The β in CIFS is set to be 2. For the ResNet-18 and its CIFS-modified version, we train models for 120 epochs with the SGD optimizer (momentum 0.9 and weight decay 0.0002). The learning rate starts from 0.1 and is multiplied with 0.1 at epoch 75 and epoch 90. For the WRN-28-10, we train model for 110 epochs with weight decay 0.0005.

In Section 4.1, we evaluate the robustness of CNNs against four white-box attacks with a perturbation budget $\epsilon = 8/255$ in l_∞ norm — FGSM, PGD-20 (step size $\epsilon/10$), C&W (optimized by PGD for 30 steps with a step size $\epsilon/10$) and PGD-100 (step size $\epsilon/10$).

Robustness Evaluation with AutoAttack: Here, we also report the robust accuracies of defense methods against AutoAttack (Croce & Hein, 2020), which consists of both white-box and black-box attacks. AutoAttack regards models to be robust at a certain data point only if the models correctly classify all types of adversarial examples generated by AutoAttack of that data point. We consider the AutoAttack including one strong white-box attack (Auto-PGD (Croce & Hein, 2020)) and one black-box attack (Square-Attack (Andriushchenko et al., 2020)). Since the Square Attack requires many queries, we sample 2000 images (200 per class) from the CIFAR10 for evaluation. The attack parameters are set according to the officially released AutoAttack[4]. From Table B.1, we observe that CIFS enjoys better robustness against AutoAttack in comparison to the vanilla ResNet-18 model and its CAS-modified version.

Best-epoch robustness during training: Due to the susceptibility of overtrained models to overfitting (Rice et al.,

[4] https://github.com/fra31/auto-attack



(a) "ship": non-CIFS (b) "ship": CIFS-sigmoid (c) "ship": CIFS-softplus (d) "ship": CIFS-softmax

(e) "autmobile": non-CIFS (f) "autmobile": CIFS-sigmoid (g) "autmobile": CIFS-softplus (h) "autmobile", CIFS-softmax

(i) "frog": non-CIFS (j) "frog": CIFS-sigmoid (k) "frog": CIFS-softplus (l) "frog", CIFS-softmax

Figure A.2. The magnitudes of channel-wise activations (top) at the penultimate layer, their activated frequency (bottom), and the weights of the last linear layer (middle) *vs.* channel indices. The robust accuracies against PGD-20 (on the whole dataset) are 46.64% for non-CIFS, 49.87% for the CIFS-sigmoid, 50.38% for the CIFS-softplus, and 51.23% for the CIFS-softmax respectively

2020), it seems reasonable to compare the results at the end of the training (and not for the best epochs) (Madry et al., 2018; Zhang et al., 2020; Rice et al., 2020). In Section 4.1, we report the robust accuracies of ResNet-18 and WRN-28-10 models at the last epochs. Here, we also provide the results at the **best** epochs for reference.

Table B.1. Robustness comparison of defense methods on CIFAR10. We report the last-epoch robust accuracies (%) against AutoAttack.

ResNet-18	Vanilla	CAS	CIFS
ResNet-18	44.00	42.70	**46.20**
WRN-28-10	47.20	46.55	**49.75**

From Table B.2, we see that, for the ResNet-18 architecture, the CIFS-modified model results in the similar best-epoch robustness (PGD-100) to that of the vanilla ResNet-18. For the WRN-28-10, the vanilla model has the better best-epoch robustness compared to the CIFS-modified version. This may be due to the fact that CIFS suppresses redundant channels and reduces the model capacity.

By comparing results in Table B.2 with those in Table 1, we observe that CIFS indeed ameliorates the overfitting of AT. Specifically, the best-epoch robust accuracy of the vanilla WRN-28-10 (resp. ResNet-18) against PGD-100 attack is 54.17% (resp.49.47%), but the last-epoch accuracy drops to 47.08% (resp. 44.72%). In contrast, for the CIFS-modified versions, the last-epoch robust accuracies against PGD-100 attack are maintained around the best-epoch ones (for WRN-28-10, from 52.03% to 51.51%; for ResNet-18, from 49.76% to 48.74%).

Table B.2. Robustness comparison of defense methods on CIFAR10. We report the best robust accuracies (%) during training. For each model, the results of the strongest attacks are marked with an underline.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	83.63	56.73	50.64	49.51	49.47
CAS	85.66	56.25	47.69	46.52	45.69
CIFS	82.46	**58.98**	**51.94**	**51.25**	**49.76**
WRN-28-10	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	86.53	**61.43**	**55.69**	**54.45**	**54.17**
CAS	87.51	58.54	52.06	51.27	50.69
CIFS	84.67	61.03	54.09	53.76	52.03

Robust accuracies for various values of β_{atk}: In Section 4.1, we evaluate the robustness of CIFS-modified models by using the adaptive loss in Equation (3). For each type of attack, we assign various values to β_{atk} and report the worst robust accuracies. Here, for reference, we provide the defense results of the ResNet-18 model on CIFAR10 for different values of β_{atk} that are used in Section 4.1. The results in Table 1 (ResNet-18) are collected from Table B.3.

B.2. Robustness Enhancement under TRADES

To improve the robustness of CNNs, various training-based strategies have been proposed, including vanilla adversarial training (AT) (Madry et al., 2018), friendly-adversarial training (FAT) (Zhang et al., 2020), and TRADES (Zhang et al., 2019). In Section 4.1, we show that CIFS can further enhance the robustness of CNNs under the vanilla AT and FAT. Here, we conduct more experiments to check whether TRADES is also suitable for CIFS.

Table B.3. Robust accuracies (%) for values of β_{atk} on CIFAR10. The value "∞" means the attack only considers the second term in Equation (3). The value "∞-1" (resp. "∞-2") means the attacker completely focuses on the first (resp. second) CIFS-modifed layer. The bracketed numbers are those reported in Table 1 (ResNet-18).

ResNet-18	β_{atk}	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	-	[84.56]	[55.11]	[46.62]	[45.95]	[44.72]
CAS	0	[86.73]	83.17	88.45	88.52	88.24
	0.1	-	58.61	61.36	85.51	62.40
	1	-	56.36	52.86	62.34	56.02
	2	-	56.06	49.76	54.94	50.62
	10	-	56.03	47.47	49.35	47.70
	100	-	56.02	47.04	48.36	46.74
	∞	-	56.02	47.06	48.31	46.55
	∞-1	-	[55.99]	[45.29]	[44.18]	[43.22]
	∞-2	-	82.68	87.87	87.79	87.72
CIFS	0	[83.86]	60.58	52.64	51.32	49.94
	0.1	-	[**58.86**]	51.40	50.88	49.42
	1	-	59.20	51.28	[**50.16**]	48.74
	2	-	59.24	[**51.23**]	50.28	48.79
	10	-	59.35	51.27	50.70	[**48.70**]
	100	-	59.38	51.41	51.04	48.80
	∞	-	59.43	51.45	51.08	48.82
	∞-1	-	61.06	54.96	53.83	52.82
	∞-2	-	60.03	52.30	50.92	50.03

Table B.4. Robustness comparison of vanilla CNNs and their CIFS-modified version under various AT-based strategies. We report the robust accuracies (%) on various types of adversarial data.

ResNet-18	Natural	FGSM	PGD-20	PGD-100
Vanilla-AT	84.56	55.11	46.62	44.72
Vanilla-TRADES	83.96	57.09	50.27	48.83
Vanilla-FAT	87.16	56.43	47.64	45.35
CIFS-AT	83.86	58.86	51.23	48.74
CIFS-TRADES	85.20	54.76	46.13	43.65
CIFS-FAT	86.35	**59.47**	**51.68**	**49.52**

From Table B.4, we observe that, for the vanilla ResNet-18 model, TRADES effectively robustifies the network and outperforms its counterparts by a large margin (e.g., 48.83% of TRADES *vs.* 44.72% of AT against PGD-100 attack). However, for the CIFS-modified models, TRADES performs worse than AT and FAT. In general, CIFS-modification in combination with the FAT training strategy achieves the best robustness against various attacks.

C. Robustness Evaluation on SVHN

Training and Evaluation details: On the SVHN dataset, we train the ResNet-18 model and its CIFS-modified version with PGD-10 adversarial examples ($\epsilon = 8/255$, step size $\epsilon/4$

with random initialization). We train models for 120 epochs with the SGD optimizer (momentum 0.9 and weight decay 0.0005). The learning rate starts from 0.01 and is multiplied with 0.1 at epoch 75 and epoch 90.

In Section 4.1, we evaluate the robustness of CNNs against four white-box attacks with a perturbation budget $\epsilon = 8/255$ in l_∞ norm — FGSM, PGD-20 (step size $\epsilon/10$), C&W (optimized by PGD for 30 steps with a step size $\epsilon/10$) and PGD-100 (step size $\epsilon/10$).

Robustness Evaluation with AutoAttack: Here, we also report the robust accuracies of defense methods against AutoAttack on SVHN (Table C.1). The evaluation settings of AutoAttack follows those in Appendix B.1.

Table C.1. Robustness comparison of defense methods on SVHN. We report the robust accuracies (%) at the last epochs.

ResNet-18	Vanilla	CAS	CIFS
AutoAttack	40.60	39.30	**42.10**

Best-epoch robustness during training: In Section 4.1, we report the robust accuracies of ResNet-18 models at the last epochs during training. Here, we report the best-epoch robustness for reference (Table C.2). We see that CIFS modified version enjoys the better best-epoch robustness in comparison to the vanilla ResNet-18 model.

Table C.2. Robustness comparison of defense methods on SVHN. We report the best robust accuracies (%) during training. For each model, the results of the strongest attack are marked with an underline.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	93.88	66.02	51.71	48.87	47.59
CAS	93.90	65.53	50.52	48.39	46.39
CIFS	93.27	**67.36**	**52.67**	**50.20**	**48.36**

D. More Results on FMNIST

Training and Evaluation details: On the FMNIST dataset, we train ResNet-10 with PGD-20 adversarial examples ($\epsilon = 0.3$, step size 0.02 with random initialization). The β in CIFS is set to be 2. We train models for 120 epochs with the SGD optimizer (momentum 0.9 and weight decay 0.0002). The learning rate starts with 0.1 and is multiplied with 0.1 at epochs 45, 75 and 90.

We evaluate the robustness of the ResNet-10 models against FGSM, PGD-20, and PGD-100 white-box attacks. The perturbation is bounded by $\epsilon = 0.3$ in l_∞ norm. The step size of PGD-20 is set to be 0.01, and that of PGD-100 is set to be 0.02. Here, we report both the last-epoch robust accuracies and the best-epoch robust accuracies in Table D.1.

Table D.1. Robustness comparison of defense methods on FMNIST. For each model, the robust accuracies (%) of the strongest attack are remarked with an underline. For each type of attack, the best defense results are highlighted in bold. Comparing the defense rates of the strongest attacks, we observe that CIFS outperforms other defenses by a large margin.

Last	Natural	FGSM	PGD-40	PGD-100
Vanilla	85.19	80.52	66.47	60.99
CAS	86.59	**82.45**	65.58	59.51
CIFS	83.35	77.48	**66.59**	**65.50**
Best	Natural	FGSM	PGD-40	PGD-100
Vanilla	85.19	81.21	67.63	63.36
CAS	86.63	**83.59**	68.73	62.65
CIFS	83.32	78.55	**69.05**	**67.21**

E. More Results on Ablation Study

E.1. Effects of β in CIFS:

Here, we train CIFS-modified ResNet-18 models on CIFAR10 with various values of β in Equation (3). The coefficient β balances the accuracies of raw predictions and the final prediction. From Table E.1, we observe that β values that are too small or too large values lead to drops in the accuracies of natural data and adversarial data. On the one hand, if the value of β is too small, the raw predictions made by CIFS are not reliable. Thus, the channels selected by CIFS may not be the truly relevant ones with respect to the ground-truth class. On the other hand, if the value of β is too large, the optimization procedure mostly considers the raw predictions, the final prediction (output) becomes unreliable. When $\beta = 2$, we achieve the best robustness against various types of attack.

Table E.1. Robustness accuracies (%) on CIFAR10 for CIFS with various values of β.

ResNet-18	Natural	FGSM	PGD-20	PGD-100
Vanilla	84.56	55.11	46.62	44.72
$\beta = 0.1$	75.22	53.41	48.10	46.28
$\beta = 1$	82.34	58.15	50.50	48.35
$\beta = 2$	83.86	**58.86**	**51.23**	**48.74**
$\beta = 10$	82.97	57.62	49.34	47.10
$\beta = 100$	75.41	52.90	45.00	43.12

E.2. Effects of the top-k feature assessment

In general, k should be larger than 1 but not too large.

If we use the top-1, once adv. data fool probe nets, the channels relevant to true labels will be missed, and this will lead to wrong predictions (Table 5, line top-1). Instead, we use top-2 for reliable channel selection. The efficacy

is attributed to *two* aspects: Firstly, the top-2 accuracies of adv. data are usually high (see Table 4), thus channels relevant to top-2 logits *include those relevant to the true class*. Secondly, Tian et al. (2021)[5] reports that CNNs' predictions of adv. data usually belong to the superclass that contains true labels. Classes (e.g., cat, dog) in the same superclass (e.g., animals) usually share similar semantic features. Thus, most of the top-2 selected channels are useful for predicting the true class.

Although the top-2 selected channels may contain info about the other wrong class, the following layers (after CIFS) are capable of "purifying" features and make better predictions. This is verified by Table 5, the results in the line top-2 (CIFS/CIFS 48.72% vs. CIFS/Final 54.96%) mean that around 6% adv. data, which successfully fool probes, are still finally correctly classified. However, too large k may degrade the relevance assessment due to too much noisy info (e.g., the effect of top-3 is worse than top-2 in Table 5).

E.3. Layers to be modified

Positions of CIFS modules: Here, we try different combinations of the layers to be modified by CIFS. In CNNs, the features of deep layers are usually more characteristic in comparison to those in the shallower layers (Zeiler & Fergus, 2014), and each channel of the features captures a distinct view of the input. The predictions often depend only on the information of a few essential views of the inputs. CIFS improves adversarial robustness by adjusting channel-wise activations. Thus, we apply CIFS to the deeper layers instead of the shallower ones. Specifically, we modify the ResNet-18 by applying CIFS at the last (P1) and/or the second last (P2) residual blocks. The experimental results are reported in Table E.2. We observe that simultaneously applying CIFS into P1 and P2 performs the best against various attacks. Intuitively, because the features can be progressively refined, applying CIFS at P1&P2 better purifies the channels compared to applying it only at P1 or P2.

Table E.2. Robustness (%) comparison of the positions where CIFS modules are placed.

ResNet-18	Natural	FGSM	PGD-20	PGD-100
Vanilla	84.56	55.11	46.62	44.72
P1	84.02	57.60	48.45	45.95
P2	82.62	56.55	47.22	44.81
P1-P2	83.86	**58.86**	**51.23**	**48.74**

E.4. Architecture of Probe Networks

Linear vs. Non-linear Probe: For a certain layer modified by CIFS, the probe network in CIFS serves as the surrogate classifier of the subsequent layers in the backbone model. Thus, the probe networks should be powerful enough to make correct predictions based on the features of this layer. For the CIFS in the last residual block, we use a linear layer network as the probe, while for the CIFS in the second last residual block, we compare the cases of using a linear layer versus using a two-layer MLP network. From Table E.3, we observe that the MLP-Linear combination shows a similar performance compared to the combination of two linear layers against adversarial attacks, but enjoys a clear advantage on the natural data (83.86% vs. 81.52%). This is because the features in the second last residual block are not as characteristic as those in the last block and cannot be linearly separated. The MLP can thus classify the features better than a pure linear layer.

Table E.3. Robustness comparison (%) of the probe architectures in CIFS modules (at P1-P2).

ResNet-18	Natural	FGSM	PGD-20	PGD-100
Vanilla	84.56	55.11	46.62	44.72
Linear-Linear	81.52	58.33	**51.32**	**49.07**
MLP-Linear	83.86	**58.86**	51.23	48.74

[5] Q. Tian, K, Kuang, F. Wu, Y. Wang, Intriguing class-wise properties of adversarial training. OpenReview. 2021