Towards Robust Training via Gradient-diversified Backpropagation

Xilin He[1], Cheng Luo[2], Qinliang Lin[1], Weicheng Xie[1,*], Muhammad Haris Khan[3],
Siyang Song[4], Linlin Shen[1]
[1]Shenzhen University, China [2]Monash University, Australia [3]MBZUAI, UAE
[4]University of Exeter, England
wcxie@szu.edu.cn

Abstract

Neural networks are prone to be vulnerable to adversarial attacks and domain shifts. Adversarial-driven methods including adversarial training and adversarial augmentation, have been frequently proposed to improve the model's robustness against adversarial attacks and distribution-shifted samples. Nonetheless, recent research on adversarial attacks has cast a spotlight on the robustness lacuna against attacks targeted at deep semantic layers. Our analysis reveals that previous adversarial-driven methods tend to generate overpowering perturbations in deep semantic layers, leading to distortion of the training for these layers. This can be primarily attributed to the exclusive utilization of loss functions on the output layer for adversarial gradient generation. This inherent practice projects an excessive adversarial impact on the deep semantic layers, elevating the difficulty of training such layers. Therefore, from the standing point of relaxing the excessive perturbations in the deep semantic layer and diversifying the adversarial gradients to ensure robust training for deep semantic layers, this paper proposes a novel Stochastic Loss Integration Method (SLIM), which can be instantiated into the existing adversarial-driven methods in a plug-and-play manner. Experimental results across diverse tasks, including classification and segmentation, as well as various areas such as adversarial robustness and domain generalization, validate the effectiveness of our proposed method. Furthermore, we provide an in-depth analysis to offer a comprehensive understanding of layer-wise training involving various loss terms.

1. Introduction

Recent advances in convolutional neural networks (CNN) have enabled remarkable success in various computer vision tasks, including classification, segmentation and object detection [11, 25, 28]. Yet CNNs are prone to be vulnerable against adversarial attacks [10, 24] and out-of-distribution (OOD) samples [12], which constrains the broader application of deep learning. Consequently, extensive studies have been dedicated to improving models' robustness against diverse input perturbations. Various defense strategies, including loss regularization [16, 19, 22], adversarial training (AT) [24, 30, 33], and data augmentation [13–15], have been proposed to defend against adversarial attacks and OOD samples.

Among the aforementioned strategies, adversarial techniques have consistently demonstrated their effectiveness. These techniques encompass adversarial training [24, 30] for countering adversarial attacks and adversarial augmentation [29, 35] to improve domain generalization. Adversarial techniques tackle a min-max optimization problem involving the loss function. In this process, they maximize the targeted loss function by introducing perturbed gradients into specific components, such as input images or feature statistics. Subsequently, they minimize this targeted optimization loss function by updating the model's parameters using gradient descent methods, which involve iteratively adjusting model parameters to approach the loss function's minimum. The majority of these methods utilize the cross-entropy (CE) loss or its variants for adversarial sample generation in training. However, recent research in adversarial attack [31] has uncovered a vulnerability in models trained using the aforementioned strategies. These models, while being effective against certain adversarial attacks, struggle to maintain robustness when confronted with attacks targeting the deep semantic layers of the network [17, 31].

To uncover the underlying reasons behind this vulnerability, we conduct a detailed comparison of module-wise traces of the Hessian matrix between a benign model and the one that has undergone adversarial training. In most cases, a lower trace of the Hessian matrix corresponds to a flatter loss landscape. A trained model with a smooth loss landscape often exhibits greater robustness and generalization capabilities [6, 18, 37]. Based on the analysis presented

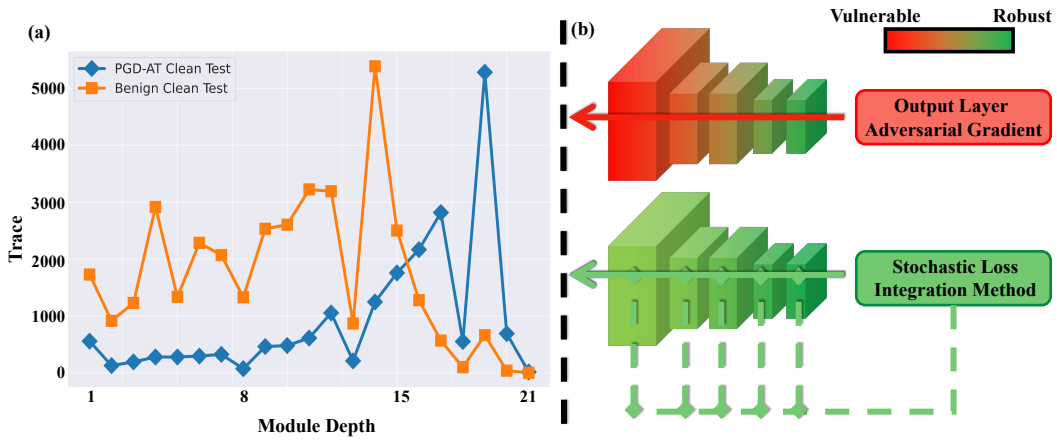

Figure 1. Illustration of (a) vulnerabilities in models trained using traditional adversarial training (PGD-AT [24]) and (b) our proposed solution towards addressing these vulnerabilities. (a) Module-wise trace of the Hessian matrix [6] computed as the second-order derivative of the loss function w.r.t. the module parameters. The trace of the Hessian matrix serves as a sensitivity measurement reflecting the module-wise flatness of the loss landscape [18, 37]. (b) Comparison between adversarial gradients backpropagated from the output layer and adversarial gradients obtained through our SLIM approach.

in Fig. 1(a), it is evident that adversarial training (i.e., PGD-AT) leads to higher traces in the deep modules of the network compared to those in a benign model. Paradoxically, while adversarial training uses output layer losses to drive the robust training, they create a robustness lacuna in the deep semantic layers. This observation could potentially serve as evidence for why adversarial learning ceases to be effective in resisting semantic layer attacks (SLA) [23, 31].

On the other hand, previous adversarial augmentation methods [29, 35, 36] mainly leverage the cross-entropy loss as the adversarial loss term to generate gradients in an augmented feature or image space that is generalized to unseen domains. However, OOD samples may result in various shifted distributions within the feature space of the deep semantic layer. Consequently, previous works have encountered limitations in achieving a comprehensive search space for adversarial augmentation to encompass various types of OOD samples.

In this paper, we undertake a comprehensive analysis of limitations observed in prior adversarial-driven methods that primarily rely on loss functions operating at the output layer. The analysis uncovers the overpowering disruptions in the deep semantic layer by adversarial training distort its training. Subsequently, in the pursuit of relaxing the overpowering noises in the deep semantic layer to ensure sufficient training for deep semantic layers, we introduce an approach aimed at *diversifying the generated adversarial gradients*, termed Stochastic Loss Integration Method (SLIM), as illustrated in Fig. 1(b), which can seamlessly integrate into various adversarial-driven methods in a plug-and-play manner to further boost the performances.

Experimental results show that despite our SLIM being simple, it records state-of-the-art performances when instantiated in adversarial training and adversarial augmentation methods. Investigations of the models' clustering effect [17] and the trace of the layer-wise Hessian matrix [6] provide further insights into how each layer of models is affected in the training process. Our contributions can be summarized as:

- We unveil that the invalidity of adversarial-driven methods against semantic layer attacks stems from the excessive distortion in the deep layers during adversarial training.
- We introduce the Stochastic Loss Integration Method (SLIM), which can be seamlessly integrated into existing adversarial-driven methods to further boost the robustness performances. SLIM acts as a relaxation mechanism for overpowering noise in deep layers, improving the overall robustness of these layers.
- Experimental results demonstrate the effectiveness of SLIM across diverse tasks and research domains. Moreover, we provide interesting insights into layer-wise training under adversarial-driven methods.

2. Related Work

2.1. Adversarial-driven Methods

Since the vulnerability of deep learning models against adversarial attacks and OOD samples has been reported [14, 24], many works studied the robustness of the models

and proposed several defense strategies. Among the various strategies proposed, one of the most effective is known as the adversarial-driven approach [24, 30, 33, 36]. This approach can be further classified into two categories: adversarial training, which focuses on defending against adversarial attacks, and adversarial augmentation, which tackles OOD samples.

Adversarial Training. Adversarial training is the most effective way of improving adversarial robustness by adapting adversarial samples for training. Mady et al. [24] propose to train the model to minimize the adversarial loss while using PGD attack to maximize it. The vast majority of adversarial training methods follow the same paradigm, but train the model with different loss objective functions to obtain better robustness performances. TRADES [33] minimizes the multi-class calibrated loss between the output of the original image and that of the adversarial examples as the surrogate loss function to substitute for cross-entropy loss. MART [30] proposes to revisit the misclassified samples and optimize the misclassification-aware regularization with the standard adversarial risk.

Adversarial Augmentation. Adversarial augmentation has recently been investigated to help models obtain stronger robustness against distribution-shifted samples. It intends to apply data augmentation under the guidance of adversarial gradients instead of randomness. Wang et al. [29] develop the adversarial variant of AugMix [13], namely as AugMax to adversarially mix multiple diverse augmented images. Their method achieves a significant improvement in OOD robustness compared to the random mixing of AugMix. Zhang et al. [35] formulate AdvStyle to explore a larger augmentation space for feature-level style augmentation with adversarially updating the statics control factors. Similar approaches have also been proposed [9, 36] for cross-domain segmentation and few-shot domain generalization.

2.2. Robustness Lacuna of Semantic Layer

Previous adversarial training methods mainly employ loss functions in the output layer solely for adversarial gradient backpropagation [24] but overlook the importance of the deep semantic layer and intermediate layers. Notably, Feature Scattering [32] generates adversarial examples by maximizing the distances between the adversarial features and the natural ones. Bai et al. [2] aim to suppress redundant channel activations in intermediate layers by adversarial examples. Recently, LAFEAT [31], an adversarial attack targeting the semantic layers, revealed that the semantic layer features can be effectively utilized for crafting the adversary, indicating the existing robustness lacuna of deep layers even for adversarial-trained models. However, we observe that the majority of the adversarial robustness works solely apply output layer attacks during training and evaluation, but do not take deep semantic layer robustness into consideration.

3. Layer-wise Adversarial Effect

Previous works in adversarial augmentation [29, 35, 36] and the majority of works in adversarial training [24, 30, 33] primarily employ output layer losses (OLL) compute backward adversarial gradients. However, relying solely on output layer losses would distort the training of the deep semantic layer with overpowering noises in these layers, resulting in a lacuna of robustness for various OOD samples and adversaries.

Adversarial Samples. Recent adversarial training methods have gained significant improvement against adversarial samples by modifying the loss functions on the output layer [24, 30, 33]. Despite gaining robustness improvement on the adversarial samples generated by adversarial gradient calculation from OLL, recent methods of adversarial training suffer from attack samples targeting deep semantic layer [31]. We reckon that this robustness lacuna of the semantic layer comes from the overpowering distortion on the semantic layer during adversarial training, resulting in sufficient training for these layers. As OLL-driven adversarial samples mainly affect features in the deep layers, solely using these adversarial samples for training leads to a robustness lacuna in the deep semantic layer.

OOD Samples. On the robustness against OOD samples, recent adversarial augmentation methods [29, 35, 36] utilize OLL losses solely to generate augmentation samples. However, since OOD samples can be corrupted in various ways, the effects specific to these samples may vary significantly across the layers of a network. To this end, adversarial augmentation is used to produce sufficient perturbations in various layers of networks. However, adversarial augmentation methods mainly perturb only shallow layers. As opposed to adversarial training, these adversarial augmentation methods have neglected the robustness of the deep layers of neural networks.

Here, we observe how diverse adversarial attacks or adversarial augmentation techniques impact various layers. To assess the influence of these different adversarial techniques, we measure cosine similarity between the features of benign samples and that of the adversarial samples across different layers to quantify the extent of the adversarial impact.

As shown in Fig. 2, solely adopting OLL to generate adversarial samples primarily leads to disruptions in the deep semantic layer. However, as shown in Fig. 1, adversarial-trained models have much higher trace values in the deep layers with adversarial test samples. This indicates that adversarial training mainly defends the model in the shallow layers, but leaves a robustness lacuna in the deep semantic layers. As adversarial training consistently creates

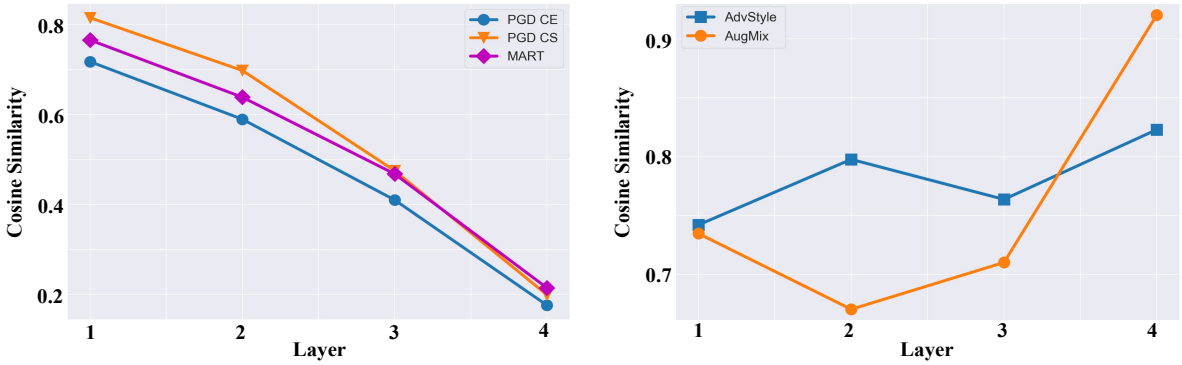

Figure 2. Cosine similarity between adversarial (PGD-CE [24], PGD-CS and MART [30]) or data augmentation (AdvStyle [35] and AugMix [13]) features and their benign features across different layers. PGD-CE and PGD-CS denote the PGD attack using cross-entropy loss and cosine similarity loss to generate adversarial samples, respectively. The cosine similarity loss is adopted for the features in the 4-th layer.

Model / Attack	PGD-20	LAFEAT	SSA
PGD-CE AT	44.66	38.07	51.21
MART	48.13	41.12	49.89
PGD-CS AT	41.28	39.70	56.84

Table 1. Recognition accuracy of adversarially-trained models under output-layer attack (PGD-20 [24]) and intermediate layers attack (LAFEAT [31] and SSA [23]). PGD-CE and PGD-CS denote the PGD training using cross-entropy loss and cosine similarity loss to generate adversarial samples, respectively. PGD-CE and MART [30] employ output-layer losses for adversarial sample generation and PGD-CS employs a semantic layer cosine similarity loss function.

overpowering perturbations in the deep semantic layer, we reckon that such excessive perturbations could distort the robust training for the deep semantic layer.

As shown in Tab. 1, semantic layer attack achieves a larger performance decline against PGD-trained models compared with PGD-20, validating the existence of robustness lacuna in the semantic layers for the adversaries. To generate different perturbations from vanilla PGD (PGD-CE) in the deep semantic space, we introduce a simple variant of PGD (PGD-CS), that employs the cosine similarity loss of the features in the deep semantic layer instead of cross-entropy loss at the output layer, to generate adversarial training samples. In this case, PGD-CS aims to minimize the cosine similarity between deep features from adversarial samples and benign samples. As shown in Tab. 1, by employing targeted perturbations on the deep semantic layer, the PGD-CS AT trained model's robustness against intermediate-layer attacks is improved. This indicates that inducing perturbations in the intermediate layers during the training can improve robustness against intermediate-layer attacks. This demonstrates that distracting the excessive perturbations generated in the deep semantic layer during adversarial training could lead to a semantic layer with stronger robustness.

Meanwhile, it can be seen that despite using PGD-CE AT could improve adversarial robustness against PGD adversarial samples, there would be a corresponding decrease in the adversarial robustness against semantic layer attack (i.e LAFEAT and SSA). This indicates that evaluations on adversarial robustness with solely PGD samples might not be comprehensive.

4. Stochastic Loss Integration Method (SLIM)

With solely leveraging output layer losses (OLL) to generate adversarial gradients providing an overpowering effect in the deep semantic layers, we propose to introduce an additional and random loss term functioned in the intermediate layers to combine with OLL to calculate adversarial gradients, namely as Stochastic Loss Integration Method (SLIM). By combining losses functioned in the intermediate layers, SLIM could relax the noises in the deep layers by affecting the calculation of the adversarial gradient, and thus prevent the overpowering noises in the deep semantic layers from distorting the training. It is worth noting that the proposed method serves as a plug-and-play strategy that can be seamlessly instantiated into methods in adversarial training and adversarial augmentation. However, for the sake of clarity, we describe how our method is inserted into adversarial training. To prevent overpowering noises generated in the deep semantic layer, the proposed SLIM mainly has two random elements: arbitrary functioned layers and stochastic loss function metric.

Arbitrary layer selection for inducing perturbation. Let f_θ^L be a neural network with $L-1$ intermediate layers and parameter θ. $x^{(l)}$ and $f_\theta^{(l)}(x)$ denote the input and output of the l- th intermediate layer. It is worth noting that all layers, excluding the final layer which outputs the probability vectors, are referred to as intermediate layers in the former definition.

To ensure a comprehensive exploration of perturbations across various intermediate layers during the training process, we randomly assign the l-th intermediate layer as the target to adversarially induce perturbation.

Diversity enhancement for intermediate layer perturbations. In order to empower the model with the ability to withstand a broad range of potential disruptions, an additional loss term functioned on the chosen intermediate layer is randomly sampled from a formulated dictionary D_{d_n} with d_n various loss function metrics, including widely used mean square error loss and cosine similarity:

$$\begin{cases} \mathcal{L}_{\text{MSE}} = \|f_\theta^l(x) - f_\theta^l(x^{adv})\|_2^2 \\ \mathcal{L}_{\text{Ortho}} = \|\frac{f_\theta^l(x) \cdot f_\theta^l(x^{adv})}{\|f_\theta^l(x)\|\|f_\theta^l(x^{adv})\|}\|_2^2 \\ \mathcal{L}_{\text{Reverse}} = -(\frac{f_\theta^l(x) \cdot f_\theta^l(x^{adv})}{\|f_\theta^l(x)\|\|f_\theta^l(x^{adv})\|}) \end{cases} \tag{1}$$

Each time we generate adversarial samples, as shown in Eq. 2, we employ both the task-specific OLL $\mathcal{L}_{\text{task}}$(e.g. cross-entropy loss) and the randomly selected intermediate layer loss function $\mathcal{L}_{\text{dict}}$ from D_{d_n}, for the computation of adversarial backward gradients. To ensure a sufficient exploration of the adversarial augmentation space, we introduce an additional mixing factor sampled from an uniform distribution, i.e., $\lambda \sim U(-1, 1)$ to control the respective influences of the two loss components in generating adversarial gradients.

$$\mathcal{L}_{adv} = \mathcal{L}_{\text{task}}(f_\theta^L(x), y) + \lambda \cdot \mathcal{L}_{\text{dict}}(f_\theta^l(x), f_\theta^l(x^{adv})) \tag{2}$$

In the context of adversarial training, as shown in Eq. 3, we iteratively update the input image x utilizing the adversarial gradients generated by the adversarial loss combination defined in Eq. 2 for t_{adv} times.

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\nabla_{x_t^{adv}} \mathcal{L}_{adv}(x_t^{adv}, y; f)) \tag{3}$$

To incorporate the proposed method into adversarial augmentation or domain adversarial training, one will only need to update the corresponding elements of the adversarial loss combination (e.g. the control factor of feature statics in adversarial feature-level style augmentation for domain generalization [35]), rather than the input images, for adversarial training.

5. Experiments

Experimental Setup. To validate the effectiveness of the proposed method, we integrate it into diverse tasks across two research domains: (i) cross-domain classification and segmentation for domain generalization and (ii) adversarial robustness evaluation for adversarial training.

Implementation Details. For domain generalization, reported results in both multi-source and single-source domain generalization are averaged over three runs. We train models for 120 epochs across all datasets. For adversarial training, we set PGD as ϵ = 8/255 and α =1/255 for PGD-AT [24]. When integrating SLIM with other methods [30, 33, 36], models were trained for 200 epochs with the same hyperparameter settings in the original literature.

For domain generalization, in both the multi-source leave-one-domain-out and single-source domain generalization scenarios, we train with the Adam optimizer. In classification tasks on PACS [21], VLCS [8], OfficeHome [27], CIFAR-10-C [12] and TerraIncognita [3], the initial learning rate and batch size are set as 2e-4 and 64, respectively. A cosine annealing schedule for adjusting the learning rate is also applied. For the cross-domain segmentation task, we follow the same hyperparameter setting with AdvStyle [35].

For adversarial training, we follow the attack setting and learning rate used in previous methods [30, 33]. The batch size is set as 64 across all the adversarial training experiments and no data augmentation techniques are included.

5.1. Domain Generalization

In the area of domain generalization, we integrate the proposed method into AdvStyle [35], which is a feature-level adversarial style augmentation method for domain generalization. To solidify the effectiveness of the proposed method, we conduct experiments on the task of domain generalized classification and segmentation. We evaluate the proposed method under both the leave-one-domain-out scenario and the more challenging single-source domain generalization scenario.

For multi-source domain generalization, under the widely adopted leave-one-domain-out setting, experiment results are shown in Tab. 2. By integrating with SLIM, AdvStyle can be further boosted across four different datasets and two network architectures. Specifically, when conducting experiments with ResNet-18 [11], performances of AdvStyle on the PACS and OfficeHome dataset can be further improved by the margins of 3.12% and 2.24%, respectively. For the more challenging single-source domain generalization scenario, experiment results are shown in Tab. 3. When integrated with the proposed SLIM, the performances of AdvStyle can still be boosted across various datasets in the two tasks of segmentation and classification, which further validates the effectiveness and versatility of the proposed SLIM.

5.2. Adversarial Training

In this section, we integrate the proposed method into widely used adversarial training methods [24, 30, 33] and evaluate the adversarial robustness against both output-layer attacks (PGD [24], AutoAttack [5]) and intermediate-layer

Method	PACS	VLCS	OfficeHome	TerraIncognita
ResNet-18				
Baseline	79.68	-	-	-
$\text{FACT}_{\text{CVPR'21}}$	84.51	-	66.56	-
$\text{StyleNeophile}_{\text{CVPR'22}}$	85.47	-	65.89	-
$\text{COMEN}_{\text{CVPR'22}}$	85.70	75.00	66.50	-
$\text{MVDG}_{\text{ECCV'22}}$	**86.56**	**77.13**	66.80	-
$\text{DSU}_{\text{ICLR'22}}$	82.70	-	66.10	-
$\text{AdvStyle}_{\text{arXiv'23, NeurIPS'22}}$	83.00	74.86	66.48	43.32
AdvStyle + SLIM	86.12	76.03	**68.72**	**45.95**
ResNet-50				
$\text{DAC-P}_{\text{CVPR'23}}$	85.60	77.00	69.50	45.80
$\text{AdvStyle}_{\text{arXiv'23, NeurIPS'22}}$	84.72	75.89	67.94	44.31
AdvStyle + SLIM	**87.03**	**77.53**	**69.21**	**46.05**

Table 2. Experiment results of multi-source domain generalization on classification under the leave-one-domain-out setting on PACS [21], VLCS [8], OfficeHome [27] and TerraIncognita [3].

Method	Segmentation (mIoU) GTA5 → Cityscapes	Classification (Acc.) PACS	Classification (Acc.) CIFAR-10-C
Baseline	37.0	46.6	74.2
pAdaIN	38.7	51.7	76.4
MixStyle	38.8	51.7	76.6
DSU	40.3	53.7	76.6
AdvStyle	41.9	58.7	78.0
AdvStyle + SLIM	**44.1**	**67.1**	**80.6**

Table 3. Experiment results of single-source domain generalization on classification and semantic segmentation. ResNet-101 (Deeplab v2), ResNet-18 and WideResNet-40-2 are adopted as the baseline settings for segmentation from GTA5 [26] to Cityscapes [4] and classification in PACS [21] and CIFAR-10-C [12], respectively.

attacks (LAFEAT [31], SSA [23]).

Tab. 4 describes the comparison of the adversarial robustness of the networks trained by various methods on the CIFAR-10 dataset [20]. As shown in Tab. 4, the performances of adversarial training methods can be further boosted by integrating with the proposed SLIM in terms of robustness against both output layer attacks and intermediate layer attacks. Specifically, the robustness of MART has been improved by 1.15% and 3.34% in terms of robustness against PGD-20 and LAFEAT, respectively. The experiment results of robustness against intermediate layer attacks further demonstrate that the proposed method can enhance the robustness of intermediate layers.

6. Analysis

In this section, we conduct an analysis of (1) the models' clustering effect [17], (2) the average trace of the module-wise Hessian matrix [6], (3) the model-wise convergence minimum [34], (4) adversarial loss over-fitting phenomenon

Method	Clean	PGD-20	AutoAttack	LAFEAT	SSA
ResNet-18	90.70	0.00	0.00	0.00	31.79
PGD-AT	84.07	46.63	38.07	38.68	51.21
TRADES	84.97	49.66	45.34	40.08	50.78
MART	81.51	55.04	41.76	41.12	49.89
AWP	80.81	54.00	-	-	-
PGD-AT + SLIM	84.96	**51.10**	38.71	43.67	63.76
TRADES + SLIM	84.71	51.37	**45.86**	43.91	**68.73**
MART + SLIM	**82.25**	55.31	42.05	**44.46**	57.21
AWP + SLIM	82.06	54.33	-	-	-

Table 4. Adversarial robustness evaluation on CIFAR-10 [20] when the proposed SLIM is integrated with widely adopted adversarial training methods, including PGD-AT [24], TRADES [33] and MART [30].

Method	Layer 1	Layer 2	Layer 3	Layer 4
Clean Test Set				
PGD-AT	32.24	39.71	57.02	77.82
PGD-AT + SLIM	34.19	45.19	61.98	77.94
AdvStlye	45.61	57.02	73.98	88.04
AdvStyle + SLIM	**48.28**	**62.32**	**80.18**	**92.16**
PGD Adversarial Test				
PGD-AT	28.42	32.31	35.66	22.26
PGD-AT + SLIM	**29.28**	**35.31**	**40.49**	**37.77**
LAFEAT Adversarial Test				
PGD-AT	26.33	29.70	32.52	28.17
PGD-AT + SLIM	**28.74**	**32.57**	**41.10**	**37.19**
OOD Corruptions Test				
AdvStyle	28.47	32.05	47.63	56.91
AdvStyle + SLIM	**30.98**	**36.65**	**51.38**	**67.69**

Table 5. Clustering accuracy of models for various test samples.

study and (5) scalability study of the loss dictionary D_{d_n}. The clustering effect provides a measurement of the model's layer-wise resistance ability against perturbations. Meanwhile, the average trace of the module-wise Hessian matrix and the model-wise parameter convergence minimum provide insights into the module-wise and model-wise smoothness of the loss landscape, respectively.

6.1. Clustering Effect

Previous work [17] introduces the clustering effect as a measurement of models' class-wise resistance ability against noises. In this section, we provide analysis from the perspective of enhancing the clustering effect to improve model robustness.

In this section, we compare the clustering effect accuracy of models trained with different methods under clean test set, adversarial attacks and OOD samples. Experiment results are shown in Tab. 5. Intuitively, a higher clustering effect accuracy indicates a stronger robustness of interme-

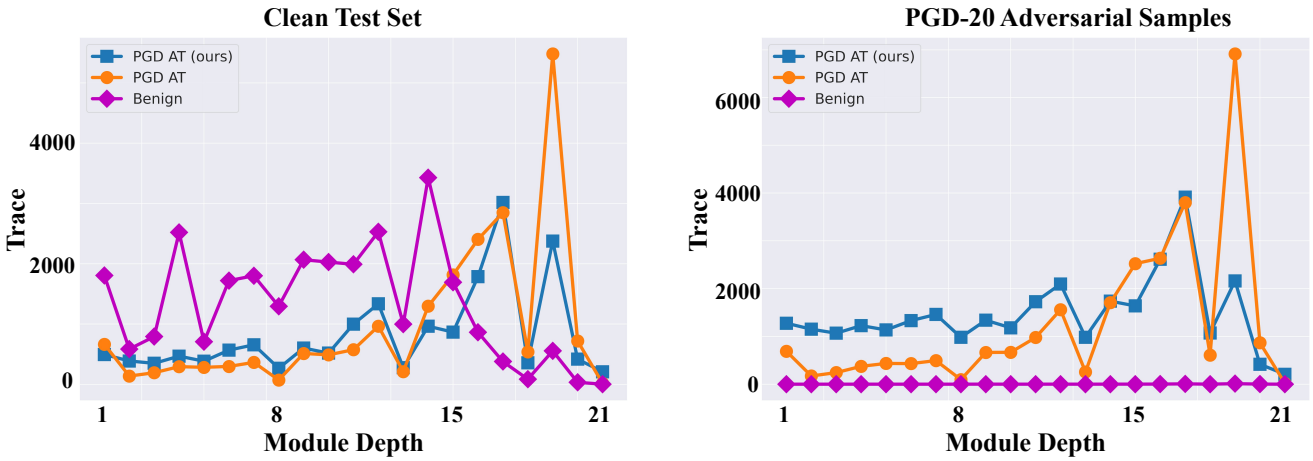

Figure 3. Trace of the module-wise Hessian matrix when testing with clean test set and PGD-20 adversarial samples. A lower trace value indicates a smoother loss landscape.

diate layers in resisting diverse perturbations. As shown in Tab. 5, models trained with the proposed method manifest a stronger clustering effect. Specifically, when undertaking intermediate-layer attacks of LAFEAT, clustering accuracy in the 3rd layer of our model significantly outperforms vanilla PGD-AT with a margin of 8.58%. As the integration of SLIM could achieve a stronger clustering effect across various methods, one can observe the effectiveness and universal applicability of the proposed method in improving layer-wise robustness.

6.2. Flatter Minimum

To study whether integrating with the proposed method can lead the convergence of the model to a flatter minimum, we follow Zhang et al. [34] to conduct the experiments of injecting Gaussian noise into the trained model weights. Fig. 5 shows the test accuracy for the cases that Gaussian noise is injected into the models' parameters. Since the PGD AT model integrated with the proposed method can resist stronger perturbations in the parameters before collapsing, we can conclude that PGD AT model integrated with SLIM achieves flatter minima in the weight space. Since it's universally acknowledged that a flatter minimum guarantees stronger robustness and generalization ability, models trained with adversarial-driven strategies integrated with the proposed method can obtain stronger robustness and generalization ability.

6.3. Average Trace of Module-wise Hessian Matrix

In this section, we investigate the average trace of the module-wise Hessian matrix of the model's parameter, which is the second-order derivative of the loss function w.r.t. the model's parameters. The average trace of the Hessian matrix can provide insights into the local geometry of the loss landscape as a sensitivity metric [6]. It is widely acknowledged that a lower trace of the model indicates a smoother local loss landscape, representing stronger robustness and generalization ability [18, 37]. Following Dong et al. [7], we compute trace information using Hutchinson's algorithm [1]. Clean test samples and adversarial samples are leveraged for calculation separately.

As shown in Fig. 3, (i) Compared with adversarial training, adversarial training protocol can obtain lower traces in the shallow layers than the naive training methods, but it suffers from a significant rise in the deep layer. (ii) In both testing scenarios, lower traces of the Hessian matrix for the deep semantic layers are obtained by PGA AT with SLIM compared with vanilla PGD AT, indicating that the generated relaxed perturbations at the deep semantic layers can indeed help these layers converge to a relatively smoother loss landscape. (iii) In the deep layer, the vanilla PGD AT model obtains much higher traces than the others, indicating that the vanilla adversarial training method mainly impacts the robustness of the shallow layers but leaving a robustness lacuna in the deep layer.

6.4. Overfitting to the Adversarial Loss

Jin et al. [17] conduct the experiments of calculating the cosine similarity between the intermediate-layer gradient versus the shift caused by the generated perturbation, to demonstrate that the adversarial gradients are overfitted to the applied output layer loss.

Following Jin et al. [17], we conduct the same experiments in Fig. 4. As shown in Fig. 4, by integrating with the proposed SLIM, numerical oscillation of cosine similarity between features shift and the adversarial gradients will occur later compared with the vanilla PGD and TRADES. This indicates that the proposed SLIM can prevent the ad-

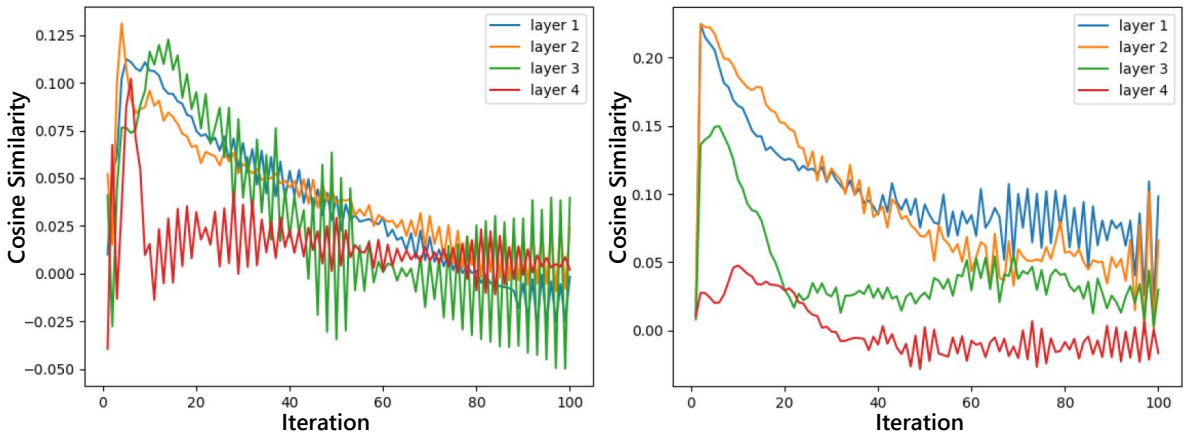

Figure 4. Layer-wise cosine similarity between the intermediate layer adversarial gradient and the feature shift caused by the generated perturbations. Experiments are conducted on the CIFAR-10 [20] with ResNet-18 [11].

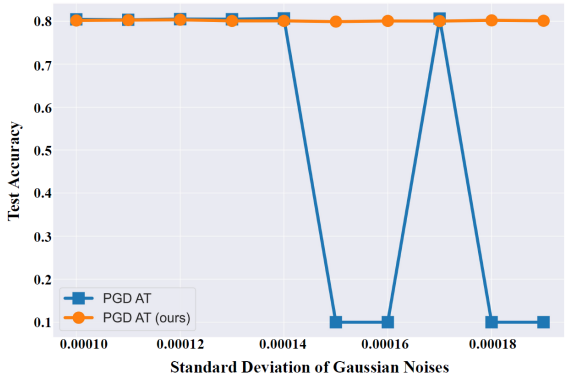

Figure 5. Comparison of test accuracy with increasing Gaussian noise injected into the models' parameters.

versarial gradients from overfitting to the applied loss combinations. Meanwhile, it can be seen that SLIM could moderate the overpowering noises in the deep layer, serving as a relaxation to the excessive distortion during adversarial training.

6.5. Scalability of Loss Dict D_{d_n}

We provide an ablation study on the formulated loss functions included in the loss dictionary D_{d_n}. For the sake of clarity, only three loss functions are included for all the experiments conducted in this paper, including MSE loss, orthogonal loss and reverse loss as shown in Eq. 1.

Since a larger loss dictionary can bring more diversified perturbations, we conduct an ablation study on the single-source domain generalization of classification on PACS to validate the high scalability of the proposed method. As shown in Tab. 6, performances on the cross-domain classification integrated with AdvStyle [35] decline when removing pre-defined loss functions in the loss dictionary, indicating that the diversity of the generated perturbations in the intermediate layers can help improve the layer-wise robustness and generalization ability. Considering the extension-friendly features of the loss dictionary, with more loss functions appended, our method may boost the performances of the existing adversarial-driven methods even further.

Loss Dictionary Setting	PACS
D_{d_n}	86.12
D_{d_n} without $\mathcal{L}_{\text{MSE}}$	85.46
D_{d_n} without $\mathcal{L}_{\text{Ortho}}$	85.39
D_{d_n} without $\mathcal{L}_{\text{Reverse}}$	85.02

Table 6. Ablation study of the formulated loss functions in the loss dictionary D_{d_n} on the PACS dataset for single-source domain generalization classification.

7. Conclusion

In this paper, we provide an analysis of the robustness lacuna of deep semantic layers from the perspective of adversarial effect, indicating the overpowering perturbations in the deep semantic layers brought by previous adversarial techniques could distort the training for these layers. Thereafter, to ensure the robustness of the deep semantic layers, we propose SLIM, the stochastic loss integration method, which can integrate into previous adversarial-driven methods of adversarial robustness and domain generalization to further boost the performances. Experiment results of domain generalization and adversarial training demonstrate the effectiveness and versatility of the proposed SLIM. Based on our SLIM, we further provide insights into layer-wise adversarial training.

References

[1] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011. 7

[2] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 472–489. Springer, 2018. 5, 6

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. 6

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020. 5

[6] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: hessian aware trace-weighted quantization of neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 6, 7

[7] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020. 7

[8] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1657–1664. IEEE Computer Society, 2013. 5, 6

[9] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24575–24584. IEEE, 2023. 3

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 5, 8

[12] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1, 5, 6

[13] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1, 3, 4

[14] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16762–16771. IEEE, 2022. 1, 2

[15] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: frequency space domain randomization for domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6891–6902. Computer Vision Foundation / IEEE, 2021. 1

[16] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[17] Yulin Jin, Xiaoyu Zhang, Jian Lou, Xu Ma, Zilong Wang, and Xiaofeng Chen. Explaining adversarial robustness of neural networks from clustering effect perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4522–4531, October 2023. 1, 2, 6, 7

[18] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1, 2, 7

[19] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive reg-

ularization for domain generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9599–9608. IEEE, 2021. 1

[20] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 6, 8

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5543–5551. IEEE Computer Society, 2017. 5, 6

[22] Dongyue Li and Hongyang R. Zhang. Improved regularization and robustness for fine-tuning in neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27249–27262, 2021. 1

[23] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15294–15303. IEEE, 2022. 2, 4, 6

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 2, 3, 4, 5, 6

[25] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022. 1

[26] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 102–118. Springer, 2016. 6

[27] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394. IEEE Computer Society, 2017. 5, 6

[28] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 1

[29] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 237–250, 2021. 1, 2, 3

[30] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1, 3, 4, 5, 6

[31] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. LAFEAT: piercing through adversarial defenses with latent features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5735–5745. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 4, 6

[32] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1829–1839, 2019. 3

[33] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019. 1, 3, 5, 6

[34] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3712–3721. IEEE, 2019. 6, 7

[35] Yabin Zhang, Bin Deng, Ruihuang Li, Kui Jia, and Lei Zhang. Adversarial style augmentation for domain generalization. *CoRR*, abs/2301.12643, 2023. 1, 2, 3, 4, 5, 8

[36] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*, 2022. 2, 3, 5

[37] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1, 2, 7